PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 :
C12N 15/10, 15/62, 15/12
C12N 15/63, C12P 21/02
C07K 13/00

A1

(11) International Publication Number: WO 92/15677

(43) International Publication Date: 17 September 1992 (17.09.92)

(21) International Application Number: PCT/US92/01456

(22) International Filing Date: 27 February 1992 (27.02.92)

(30) Priority data:
664,989 1 March 1991 (01.03.91) US

(71) Applicant: PROTEIN ENGINEERING CORPORATION [US/US]; 765 Concord Avenue, Cambridge, MA 02138 (US).

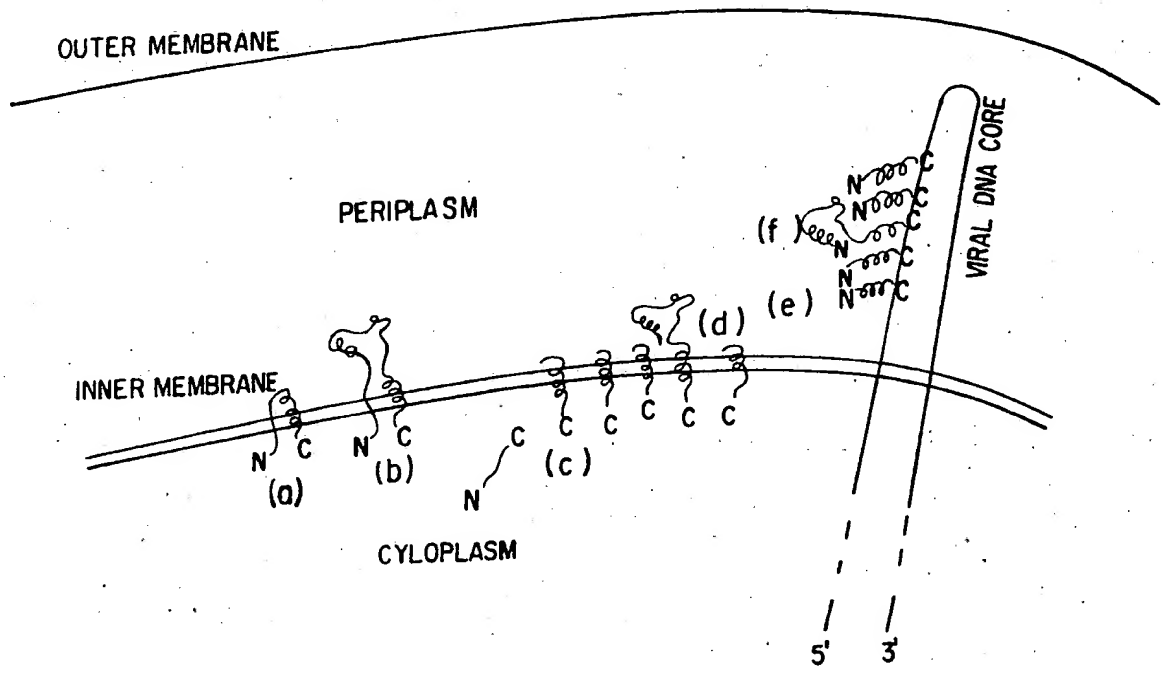
(72) Inventors: LADNER, Robert, Charles ; 3827 Green Valley Road, Ijamsville, MD 21754 (US). ROBERTS, Bruce, Lindsay ; 26 Windsor Road, Milford, MA 01757 (US). LEY, Arthur, Charles ; 122 Adena Road, Newton, MA 02165 (US). KENT, Rachel, Baribault ; 60 Stonehedge Place, Boxborough, MA 01719 (US).

(74) Agent: COOPER, Iver, P.; Browdy & Neimark, 419 Seventh Street, Ste. 300, N.W., Washington, DC 20004 (US).

(81) Designated States: AT (European patent), AU, BE (European patent), CA, CH (European patent), DE (European patent), DK, DK (European patent), ES (European patent), FI, FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, LU (European patent), MC (European patent), NL (European patent), SE (European patent).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: PROCESS FOR THE DEVELOPMENT OF BINDING MINI-PROTEINS



(57) Abstract

Genes encoding disulfide-bonded micro-proteins and metal ion-coordinated mini-proteins are semi-randomly mutagenized and expressed in suitable cells so as to result in the simultaneous display of the different mutant proteins on the surfaces of bacterial cells, spores or phage; the resulting display library is screened for members having the ability to bind to a target of interest.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	**FI**	Finland	**ML**	Mali
AU	Australia	**FR**	France	**MN**	Mongolia
BB	Barbados	**GA**	Gabon	**MR**	Mauritania
BE	Belgium	**GB**	United Kingdom	**MW**	Malawi
BF	Burkina Faso	**GN**	Guinea	**NL**	Netherlands
BG	Bulgaria	**GR**	Greece	**NO**	Norway
BJ	Benin	**HU**	Hungary	**PL**	Poland
BR	Brazil	**IE**	Ireland	**RO**	Romania
CA	Canada	**IT**	Italy	**RU**	Russian Federation
CF	Central African Republic	**JP**	Japan	**SD**	Sudan
CG	Congo	**KP**	Democratic People's Republic of Korea	**SE**	Sweden
CH	Switzerland			**SN**	Senegal
CI	Côte d'Ivoire	**KR**	Republic of Korea	**SU**	Soviet Union
CM	Cameroon	**LI**	Liechtenstein	**TD**	Chad
CS	Czechoslovakia	**LK**	Sri Lanka	**TG**	Togo
DE	Germany	**LU**	Luxembourg	**US**	United States of America
DK	Denmark	**MC**	Monaco		
ES	Spain	**MG**	Madagascar		

PROCESS FOR THE DEVELOPMENT OF BINDING MINI-PROTEINS

BACKGROUND OF THE INVENTION

Field of the Invention

This invention relates to development of novel binding mini-proteins, and especially micro-proteins, by an iterative process of mutagenesis, expression, affinity selection, and amplification. In this process, a gene encoding a mini-protein potential binding domain, said gene being obtained by random mutagenesis of a limited number of predetermined codons, is fused to a genetic element which causes the resulting chimeric expression product to be displayed on the outer surface of a virus (especially a filamentous phage) or a cell. Affinity selection is then used to identify viruses or cells whose genome includes such a fused gene which coded for the protein which bound to the chromatographic target.

Description of the Related Art

The amino acid sequence of a protein determines its three-dimensional (3D) structure, which in turn determines protein function. Some residues on the polypeptide chain are more important than others in determining the 3D structure of a protein, and hence its ability to bind, non-covalently, but very tightly and specifically, to characteristic target molecules.

"Protein engineering" is the art of manipulating the sequence of a protein in order, e.g., to alter its binding characteristics. The factors affecting protein binding are known, but designing new complementary surfaces has proved difficult. Quiocho et al. (QUIO87) suggest it is unlikely that, using current protein engineering methods, proteins can be constructed with binding properties superior to those of proteins that occur naturally.

Nonetheless, there have been some isolated successes. For example, Wilkinson et al. (WILK84) reported that a

mutant of the tyrosyl tRNA synthetase of Bacillus stearothermophilus with the mutation Thr_{51}-->Pro exhibits a 100-fold increase in affinity for ATP.

With the development of recombinant DNA techniques, it became possible to obtain a mutant protein by mutating the gene encoding the native protein and then expressing the mutated gene. Several mutagenesis strategies are known. One, "protein surgery" (DILL87), involves the introduction of one or more predetermined mutations within the gene of choice. A single polypeptide of completely predetermined sequence is expressed, and its binding characteristics are evaluated.

At the other extreme is random mutagenesis by means of relatively nonspecific mutagens such as radiation and various chemical agents. See Ho et al. (HOCJ85) and Lehtovaara, EP Appln. 285,123.

It is possible to randomly vary predetermined nucleotides using a mixture of bases in the appropriate cycles of a nucleic acid synthesis procedure. (OLIP86, OLIP87) The proportion of bases in the mixture, for each position of a codon, will determine the frequency at which each amino acid will occur in the polypeptides expressed from the degenerate DNA population. (REID88a; VERS86a; VERS86b). The problem of unequal abundance of DNA encoding different amino acids is not discussed.

Ferenci and collaborators have published a series of papers on the chromatographic isolation of mutants of the maltose-transport protein LamB of E. coli (FERE82a, FERE82b, FERE83, FERE84, CLUN84, HEIN87 and papers cited therein). The mutants were either spontaneous or induced with nonspecific chemical mutagens. Levels of mutagenesis were picked to provide single point mutations or single insertions of two residues. No multiple mutations were sought or found.

While variation was seen in the degree of affinity for the conventional LamB substrates maltose and starch, there

was no selection for affinity to a target molecule not bound at all by native LamB, and no multiple mutations were sought or found. FERE84 speculated that the affinity chromatographic selection technique could be adapted to development of similar mutants of other "important bacterial surface-located enzymes", and to selecting for mutations which result in the relocation of an intracellular bacterial protein to the cell surface. Ferenci's mutant surface proteins would not, however, have been chimeras of a bacterial surface protein and an exogenous or heterologous binding domain.

Ferenci also taught that there was no need to clone the structural gene, or to know the protein structure, active site, or sequence. The method of the present invention, however, specifically utilizes a cloned structural gene. It is not possible to construct and express a chimeric, outer surface-directed potential binding protein-encoding gene without cloning.

Ferenci did not limit the mutations to particular loci Substitutions were limited by the nature of the mutagen rather than by the desirability of particular amino acid types at a particular site. In the present invention, knowledge of the protein structure, active site and/or sequence is used as appropriate to predict which residues are most likely to affect binding activity without unduly destabilizing the protein, and the mutagenesis is focused upon those sites. Ferenci does not suggest that surface residues should be preferentially varied. In consequence, Ferenci's selection system is much less efficient than that disclosed herein.

A number of researchers have directed unmutated foreign antigenic epitopes to the surface of bacteria or phage, fused to a native bacterial or phage surface protein, and demonstrated that the epitopes were recognized by antibodies. Thus, Charbit, et al. (CHAR86a,b) genetically inserted

the C3 epitope of the VP1 coat protein of poliovirus into the LamB outer membrane protein of E. coli, and determined immunologically that the C3 epitope was exposed on the bacterial cell surface. Charbit, et al. (CHAR87) likewise produced chimeras of LamB and the A (or B) epitopes of the preS2 region of hepatitis B virus.

A chimeric LacZ/OmpB protein has been expressed in E. coli and is, depending on the fusion, directed to either the outer membrane or the periplasm (SILH77). A chimeric LacZ/OmpA surface protein has also been expressed and displayed on the surface of E. coli cells (WEIN83). Others have expressed and displayed on the surface of a cell chimeras of other bacterial surface proteins, such as E. coli type 1 fimbriae (HEDE89) and Bacterioides nodusus type 1 fimbriae (JENN89). In none of the recited cases was the inserted genetic material mutagenized.

Dulbecco (DULB86) suggests a procedure for incorporating a foreign antigenic epitope into a viral surface protein so that the expressed chimeric protein is displayed on the surface of the virus in a manner such that the foreign epitope is accessible to antibody. In 1985 Smith (SMIT85) reported inserting a nonfunctional segment of the EcoRI endonuclease gene into gene III of bacteriophage f1, "in phase". The gene III protein is a minor coat protein necessary for infectivity. Smith demonstrated that the recombinant phage were adsorbed by immobilized antibody raised against the EcoRI endonuclease, and could be eluted with acid. De la Cruz et al. (DELA88) have expressed a fragment of the repeat region of the circumsporozoite protein from Plasmodium falciparum on the surface of M13 as an insert in the gene III protein. They showed that the recombinant phage were both antigenic and immunogenic in rabbits, and that such recombinant phage could be used for B epitope mapping. The researchers suggest that similar

recombinant phage could be used for T epitope mapping and for vaccine development.

None of these researchers suggested mutagenesis of the inserted material, nor is the inserted material a complete binding domain conferring on the chimeric protein the ability to bind specifically to a receptor other than the antigen combining site of an antibody.

McCafferty et al. (MCCA90) expressed a fusion of an Fv fragment of an antibody to the N-terminal of the pIII protein. The Fv fragment was not mutated.

Parmley and Smith (PARM88) suggested that an epitope library that exhibits all possible hexapeptides could be constructed and used to isolate epitopes that bind to antibodies. In discussing the epitope library, the authors did not suggest that it was desirable to balance the representation of different amino acids. Nor did they teach that the insert should encode a complete domain of the exogenous protein. Epitopes are considered to be unstructured peptides as opposed to structured proteins.

Scott and Smith (SCOT90) and Cwirla et al. (CWIR90) prepared "epitope libraries" in which potential hexapeptide epitopes for a target antibody were randomly mutated by fusing degenerate oligonucleotides, encoding the epitopes, with gene III of fd phage, and expressing the fused gene in phage-infected cells. The cells manufactured fusion phage which displayed the epitopes on their surface; the phage which bound to immobilized antibody were eluted with acid and studied. In both cases, the fused gene featured a segment encoding a spacer region to separate the variable region from the wild type pIII sequence so that the varied amino acids would not be constrained by the nearby pIII sequence. Devlin et al. (DEVL90) similarly screened, using M13 phage, for random 15 residue epitopes recognized by streptavidin. Again, a spacer was used to move the random peptides away from the rest of the chimeric phage protein.

These references therefore taught away from constraining the conformational repertoire of the mutated residues.

Another problem with the Scott and Smith, Cwirla et al., and Devlin et al., libraries was that they provided a highly biased sampling of the possible amino acids at each position. Their primary concern in designing the degenerate oligonucleotide encoding their variable region was to ensure that all twenty amino acids were encodible at each position; a secondary consideration was minimizing the frequency of occurrence of stop signals. Consequently, Scott and Smith and Cwirla et al. employed NNK (N=equal mixture of G, A, T, C; K=equal mixture of G and T) while Devlin et al. used NNS (S=equal mixture of G and C). There was no attempt to minimize the frequency ratio of most favored-to-least favored amino acid, or to equalize the rate of occurrence of acidic and basic amino acids.

Devlin et al. characterized several affinity- selected streptavidin-binding peptides, but did not measure the affinity constants for these peptides. Cwirla et al. did determine the affinity constant for his peptides, but were disappointed to find that his best hexapeptides had affinities (350-300nM), "orders of magnitude" weaker than that of the native Met-enkephalin epitope (7nM) recognized by the target antibody. Cwirla et al. speculated that phage bearing peptides with higher affinities remained bound under acidic elution, possibly because of multivalent interactions between phage (carrying about 4 copies of pIII) and the divalent target IgG. Scott and Smith were able to find peptides whose affinity for the target antibody (A2) was comparable to that of the reference myohemerythrin epitope (50nM). However, Scott and Smith likewise expressed concern that some high-affinity peptides were lost, possibly through irreversible binding of fusion phage to target.

Lam, et al. (LAM91) created a pentapeptide library by nonbiological synthesis on solid supports. While they teach

that it is desirable to obtain the universe of possible random pentapeptides in roughly equimolar proportions, they deliberately excluded cysteine, to eliminate any possibility of disulfide crosslinking.

Ladner, Glick, and Bird, WO88/06630 (publ. 7 Sept. 1988 and having priority from US application 07/021,046, assigned to Genex Corp.) (LGB) speculate that diverse single chain antibody domains (SCAD) may be screened for binding to a particular antigen by varying the DNA encoding the combining determining regions of a single chain antibody, subcloning the SCAD gene into the gpV gene of phage λ so that a SCAD/gpV chimera is displayed on the outer surface of phage λ, and selecting phage which bind to the antigen through affinity chromatography. The only antigen mentioned is bovine growth hormone. No other binding molecules, targets, carrier organisms, or outer surface proteins are discussed. Nor is there any mention of the method or degree of mutagenesis. Furthermore, there is no teaching as to the exact structure of the fusion nor of how to identify a successful fusion or how to proceed if the SCAD is not displayed.

Ladner and Bird, WO88/06601 (publ. 7 September 1988) suggest that single chain "pseudodimeric" repressors (DNA-binding proteins) may be prepared by mutating a putative linker peptide followed by <u>in vivo</u> selection that mutation and selection may be used to create a dictionary of recognition elements for use in the design of asymmetric repressors. The repressors are not displayed on the outer surface of an organism.

Methods of identifying residues in protein which can be replaced with a cysteine in order to promote the formation of a protein-stabilizing disulfide bond are given in Pantoliano and Ladner, U.S. Patent No. 4,903,773 (PANT90), Pantoliano and Ladner (PANT87), Pabo and Suchenek (PABO86), MATS89, and SAUE86.

Ladner, et al., W090/02809 describes semirandom mutagenesis ("variegation") of known proteins displayed as domains of semiartificial outer surface proteins of bacteria, phage or spores, and affinity selection of mutants having desired binding characteristics. The smallest proteins specifically mentioned in W090/02809 are crambin (3:40, 4:32, 16:26 disulfides; 46 AAs), the third domain of ovomucoid (8:38, 16:35 and 24:56 disulfides; 56 AAs), and BPTI (5:55, 14:38, 30:51 disulfides; 58 AAs). W090/02809 also specifically describes a strategy for "variegating" a codon to obtain a mix of all twenty amino acids at that position in approximately equal proportions.

Bass, et al. (BASS90) fused human growth hormone to the gene III protein of M13 phage. He suggested that hGH and other "large proteins" might be mutated and "binding selections" applied.

SUMMARY OF THE INVENTION

A polypeptide is a polymer composed of a single chain of the same or different amino acids joined by peptide bonds. Linear peptides can take up a very large number of different conformations through internal rotations about the main chain single bonds of each α carbon. These rotations are hindered to varying degrees by side groups, with glycine interfering the least, and valine, isoleucine and, especially, proline, the most. A polypeptide of 20 residues may have 10^{20} different conformations which it may assume by various internal rotations.

Proteins are polypeptides which, as a result of stabilizing interactions between amino acids that are not necessarily in adjacent positions in the chain, have folded into a well-defined conformation. This folding is usually essential to their biological activity.

For polypeptides of 40-60 residues or longer, noncovalent forces such as hydrogen bonds, salt bridges, and

hydrophobic interactions are sufficient to stabilize a particular folding or conformation. The polypeptide's constituent segments are held to more or less that conformation unless it is perturbed by a denaturant such as high temperature, or low or high pH, whereupon the polypeptide unfolds or "melts". The smaller the peptide, the more likely it is that its conformation will be determined by the environment. If a small unconstrained peptide has biological activity, the peptide ligand will be in essence a random coil until it comes into proximity with its receptor. The receptor accepts the peptide only in one or a few conformations because alternative conformations are disfavored by unfavorable van der Waals and other non-covalent interactions.

Small polypeptides have potential advantages over larger polypeptides when used as therapeutic or diagnostic agents, including (but not limited to):

a) better penetration into tissues,
b) faster elimination from the circulation (important for imaging agents),
c) lower antigenicity, and
d) higher activity per mass.

Moreover, polypeptides, especially those of less than about 40 residues, have the advantage of accessibility via chemical synthesis; polypeptides of under about 30 residues are particularly preferred. Thus, it would be desirable to be able to employ the combination of mutation and affinity selection to identify small polypeptides which bind a target of choice.

Most polypeptides of this size, however, have disadvantages as binding molecules. According to Olivera et al. (OLIV90a): "Peptides in this size range normally equilibrate among many conformations (in order to have a fixed conformation, proteins generally have to be much larger)." Specific binding of a peptide to a target molecule requires

the peptide to take up one conformation that is complementary to the binding site. For a decapeptide with three isoenergetic conformations (e.g., β strand, α helix, and reverse turn) at each residue, there are about $6 \cdot 10^4$ possible overall conformations. Assuming these conformations to be equi-probable for the unconstrained decapeptide, if only one of the possible conformations bound to the binding site, then the affinity of the peptide for the target would be expected to be about $6 \cdot 10^4$ higher if it could be constrained to that single effective conformation. Thus, the unconstrained decapeptide, relative to a decapeptide constrained to the correct conformation, would be expected to exhibit lower affinity. It would also exhibit lower specificity, since one of the other conformations of the unconstrained decapeptide might be one which bound tightly to a material other than the intended target. By way of corollary, it could have less resistance to degradation by proteases, since it would be more likely to provide a binding site for the protease.

The present invention overcomes these problems, while retaining the advantages of smaller polypeptides, by identifying novel mini-proteins having the desired binding characteristics. Mini-Proteins are small polypeptides which, while too small to have a stable conformation as a result of noncovalent forces alone, are covalently crosslinked (e.g., by disulfide bonds) into a stable conformation and hence have biological activities more typical of larger protein molecules than of unconstrained polypeptides of comparable size. THe mini-proteins with which the present invention is particularly concerned fall into two categories: (a) disulfide-bonded micro-proteins of less than 40 amino acids; and (b) metal ion-coordinated mini-proteins of less than 60 amino acids.

The present invention relates to the construction, expression, and selection of mutated genes that specify novel mini-proteins with desirable binding properties, as well as these mini-proteins themselves, and the "libraries" of mutant "genetic packages" used to display the mini-proteins to a potential "target" material. The "targets" may be, but need not be, proteins. Targets may include other biological or synthetic macromolecules as well as other organic and inorganic substances.

The prior application, WO90/02809 generally teaches that stable protein domains may be mutated in order to identify new proteins with desirable binding characteristics. Among the suitable "parental" proteins which it specifically identifies as useful for this purpose are three proteins--BPTI (58 residues), the third domain of ovomucoid (56 residues), and crambin (46 residues)--which are in the size range of 40-60 residues wherein noncovalent interactions between nonadjacent amino acids become significant; all three also contain three disulfide bonds that enhance the stability of the molecule.

Nowhere in WO90/02809 does one find any specific recognition that a polypeptide with less than 40 residues, and especially those with only one or two disulfide bonds, would have sufficient stability to serve as a "scaffolding" for mutational variation. These "micro-proteins" are, nonetheless, of great utility, as previously indicated.

WO90/02809 also suggests the use of a protein, azurin, having a different form of crosslink (Cu:CYS,HIS,HIS,MET). However, azurin has 128 amino acids, so it cannot possibly be considered a mini-protein. The present invention relates to the use of mini-proteins of less than 60 amino acids which feature a metal ion-coordinated crosslink.

By virtue of the present invention, proteins are obtained which can bind specifically to targets other than the antigen-combining sites of antibodies. A protein is not

to be considered a "binding protein" merely because it can be bound by an antibody (see definition of "binding protein" which follows). While almost any amino acid sequence of more than about 6-8 amino acids is likely, when linked to an immunogenic carrier, to elicit an immune response, any given random polypeptide is unlikely to satisfy the stringent definition of "binding protein" with respect to minimum affinity and specificity for its substrate. It is only by testing numerous random polypeptides simultaneously (and, in the usual case, controlling the extent and character of the sequence variation, i.e., limiting it to residues of a potential binding domain having a stable structure, the residues being chosen as more likely to affect binding than stability) that this obstacle is overcome.

The appended claims are hereby incorporated by reference into this specification as an enumeration of the preferred embodiments.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows the main chain of scorpion toxin (Brookhaven Protein Data Bank entry 1SN3) residues 20 through 42. CYS_{25} and CYS_{41} are shown forming a disulfide. In the native protein these groups form disulfides to other cysteines, but no main-chain motion is required to bring the gamma sulphurs into acceptable geometry. Residues, other than GLY, are labeled at the β carbon with the one-letter code.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

I. INTRODUCTION

The fundamental principle of the invention is one of forced evolution. In nature, evolution results from the combination of genetic variation, selection for advantageous traits, and reproduction of the selected individuals,

thereby enriching the population for the trait. The present invention achieves genetic variation through controlled random mutagenesis ("variegation") of DNA, yielding a mixture of DNA molecules encoding different but related potential binding domains that are mutants of micro-proteins. It selects for mutated genes that specify novel proteins with desirable binding properties by 1) arranging that the product of each mutated gene be displayed on the outer surface of a replicable genetic package (GP) (a cell, spore or virus) that contains the gene, and 2) using affinity selection -- selection for binding to the target material -- to enrich the population of packages for those packages containing genes specifying proteins with improved binding to that target material. Finally, enrichment is achieved by allowing only the genetic packages which, by virtue of the displayed protein, bound to the target, to reproduce. The evolution is "forced" in that selection is for the target material provided and in that particular codons are mutagenized at higher-than-natural frequencies.

The display strategy is first perfected by modifying a genetic package to display a stable, structured domain (the "initial potential binding domain", IPBD) for which an affinity molecule (which may be an antibody) is obtainable. The success of the modifications is readily measured by, e.g., determining whether the modified genetic package binds to the affinity molecule.

The IPBD is chosen with a view to its tolerance for extensive mutagenesis. Once it is known that the IPBD can be displayed on a surface of a package and subjected to affinity selection, the gene encoding the IPBD is subjected to a special pattern of multiple mutagenesis, here termed "variegation", which after appropriate cloning and amplification steps leads to the production of a population of genetic packages each of which displays a single potential binding domain (a mutant of the IPBD), but which collective-

ly display a multitude of different though structurally related potential binding domains (PBDs). Each genetic package carries the version of the pbd gene that encodes the PBD displayed on the surface of that particular package. Affinity selection is then used to identify the genetic packages bearing the PBDs with the desired binding characteristics, and these genetic packages may then be amplified. After one or more cycles of enrichment by affinity selection and amplification, the DNA encoding the successful binding domains (SBDs) may then be recovered from selected packages.

If need be, the DNA from the SBD-bearing packages may then be further "variegated", using an SBD of the last round of variegation as the "parental potential binding domain" (PPBD) to the next generation of PBDs, and the process continued until the worker in the art is satisfied with the result. Because of the structural and evolutionary relationship between the IPBD and the first generation of PBDs, the IPBD is also considered a "parental potential binding domain" (PPBD).

When micro-proteins are variegated, the residues which are covalently crosslinked in the parental molecule are left unchanged, thereby stabilizing the conformation. For example, in the variegation of a disulfide bonded micro-protein, certain cysteines are invariant so that under the conditions of expression and display, covalent crosslinks (e.g., disulfide bonds between one or more pairs of cysteines) form, and substantially constrain the conformation which may be adopted by the hypervariable linearly intermediate amino acids. In other words, a constraining scaffolding is engineered into polypeptides which are otherwise extensively randomized.

Once a micro-protein of desired binding characteristics is characterized, it may be produced, not only by recombinant DNA techniques, but also by nonbiological synthetic methods.

For the purposes of the appended claims, a protein P is a "binding protein" if for at least one molecular, ionic or atomic species A, other than the variable domain of an antibody, the dissociation constant K_D (P,A) < 10^{-6} moles/liter (preferably, < 10^{-7} moles/liter).

The exclusion of "variable domain of an antibody" in (1) above is intended to make clear that for the purposes herein a protein is not to be considered a "binding protein" merely because it is antigenic.

Most larger proteins fold into distinguishable globules called domains (ROSS81). Protein domains have been defined various ways; definitions of "domain" which emphasize stability -- retention of the overall structure in the face of perturbing forces such as elevated temperatures or chaotropic agents -- are favored, though atomic coordinates and protein sequence homology are not completely ignored.

When a domain of a protein is primarily responsible for the protein's ability to specifically bind a chosen target, it is referred to herein as a "binding domain" (BD).

The term "variegated DNA" (vgDNA) refers to a mixture of DNA molecules of the same or similar length which, when aligned, vary at some codons so as to encode at each such codon a plurality of different amino acids, but which encode only a single amino acid at other codon positions. It is further understood that in variegated DNA, the codons which are variable, and the range and frequency of occurrence of the different amino acids which a given variable codon encodes, are determined in advance by the synthesizer of the DNA, even though the synthetic method does not allow one to know, a priori, the sequence of any individual DNA molecule in the mixture. The number of designated variable codons in the variegated DNA is preferably no more than 20 codons, and more preferably no more than 5-10 codons. The mix of amino

acids encoded at each variable codon may differ from codon to codon.

A population of genetic packages into which variegated DNA has been introduced is likewise said to be "variegated".

For the purposes of this invention, the term "potential binding protein" (PBP) refers to a protein encoded by one species of DNA molecule in a population of variegated DNA wherein the region of variation appears in one or more subsequences encoding one or more segments of the polypeptide having the potential of serving as a binding domain for the target substance.

A "chimeric protein" is a fusion of a first amino acid sequence (protein) with a second amino acid sequence defining a domain foreign to and not substantially homologous with any domain of the first protein. A chimeric protein may present a foreign domain which is found (albeit in a different protein) in an organism which also expresses the first protein, or it may be an "interspecies", "intergeneric", etc. fusion of protein structures expressed by different kinds of organisms.

One amino acid sequence of the chimeric proteins of the present invention is typically derived from an outer surface protein of a "genetic package" (GP) as hereafter defined. One which displays a PBD on its surface is a GP (PBD). The second amino acid sequence is one which, if expressed alone, would have the characteristics of a protein (or a domain thereof) but is incorporated into the chimeric protein as a recognizable domain thereof. It may appear at the amino or carboxy terminal of the first amino acid sequence (with or without an intervening spacer), or it may interrupt the first amino acid sequence. The first amino acid sequence may correspond exactly to a surface protein of the genetic package, or it may be modified, e.g., to facilitate the display of the binding domain.

II. MICRO- AND OTHER MINI-PROTEINS

In the present invention, disulfide bonded micro-proteins and metal-containing mini-proteins are used both as IPBDs in verifying a display strategy, and as PPBDs in actually seeking to obtain a BD with the desired target-binding characteristics. Unless otherwise stated or required by context, references herein to IPBDs should be taken to apply, mutatis mutandis, to PPBDs as well.

For the purpose of the appended claims, a micro-protein has between about six and about forty residues; micro-proteins are a subset of mini-proteins, which have less than about sixty residues. Since micro-proteins form a subset of mini-proteins, for convenience the term mini-proteins will be used on occasion to refer to both disulfide-bonded micro-proteins and metal-coordinated mini-proteins.

The IPBD may be a mini-protein with a known binding activity, or one which, while not possessing a known binding activity, possesses a secondary or higher structure that lends itself to binding activity (clefts, grooves, etc.). When the IPBD does have a known binding activity, it need not have any specific affinity for the target material. The IPBD need not be identical in sequence to a naturally-occurring mini-protein; it may be a "homologue" with an amino acid sequence which "substantially corresponds" to that of a known mini-protein, or it may be wholly artificial.

In determining whether sequences should be deemed to "substantially correspond", one should consider the following issues: the degree of sequence similarity when the sequences are aligned for best fit according to standard algorithms, the similarity in the connectivity patterns of any crosslinks (e.g., disulfide bonds), the degree to which the proteins have similar three-dimensional structures, as indicated by, e.g., X-ray diffraction analysis or NMR, and the degree to which the sequenced proteins have similar

biological activity. In this context, it should be noted that among the serine protease inhibitors, there are families of proteins recognized to be homologous in which there are pairs of members with as little as 30% sequence homology.

A candidate IPBD should meet the following criteria:

1) a domain exists that will remain stable under the conditions of its intended use (the domain may comprise the entire protein that will be inserted, e.g. α-conotoxin GI (OLIV90a), or CMTI-III (MCWH89),
2) knowledge of the amino acid sequence is obtainable, and
3) a molecule is obtainable having specific and high affinity for the IPBD, abbreviated AfM(IPBD).

If only one species of molecule having affinity for IPBD (AfM(IPBD)) is available, it will be used to: a) detect the IPBD on the GP surface, b) optimize expression level and density of the affinity molecule on the matrix, and c) determine the efficiency and sensitivity of the affinity separation. One would prefer to have available two species of AfM(IPBD), one with high and one with moderate affinity for the IPBD. The species with high affinity would be used in initial detection and in determining efficiency and sensitivity, and the species with moderate affinity would be used in optimization.

If the IPBD is not itself a known binding protein, or if its native target has not been purified, an antibody raised against the IPBD may be used as the affinity molecule. Use of an antibody for this purpose should not be taken to mean that the antibody is the ultimate target.

There are many candidate IPBDs for which all of the above information is available or is reasonably practical to obtain, for example, CMTI-III (29 residues) (CMTI-type inhibiters are described in OTLE87, FAVE89, WIEC85, MCWH89, BODE89, HOLA89a,b), heat-stable enterotoxin (ST-Ia of E.

coli) (18 residues) (GUAR89, BHAT86, SEKI85, SHIM87, TAKA85, TAKE90, THOM85a,b, YOSH85, DALL90, DWAR89, GARI87, GUZM89, GUZM90, HOUG84, KUBO89, KUPE90, OKAM87, OKAM88, AND OKAM90), α-Conotoxin GI (13 residues) (HASH85, ALMQ89), μ-Conotoxin GIII (22 residues) (HIDO90), and Conus King Kong micro-protein (27 residues) (WOOD90). Structural information can be obtained from X-ray or neutron diffraction studies, NMR, chemical cross linking or labeling, modeling from known structures of related proteins, or from theoretical calculations. 3D structural information obtained by X-ray diffraction, neutron diffraction or NMR is preferred because these methods allow localization of almost all of the atoms to within defined limits. Table 50 lists several preferred IPBDs.

Mutations may reduce the stability of the PBD. Hence the chosen IPBD should preferably have a high melting temperature, e.g., at least 50°C, and preferably be stable over a wide pH range, e.g., 8.0 to 3.0, but more preferably 11.0 to 2.0, so that the SBDs derived from the chosen IPBD by mutation and selection-through-binding will retain sufficient stability. Preferably, the substitutions in the IPBD yielding the various PBDs do not reduce the melting point of the domain below ~40°C. It will be appreciated that mini-proteins contain covalent crosslinks, such as one or more disulfides, are therefore are likely to be sufficiently stable.

In vitro, disulfide bridges can form spontaneously in polypeptides as a result of air oxidation. Matters are more complicated in vivo. Very few intracellular proteins have disulfide bridges, probably because a strong reducing environment is maintained by the glutathione system. Disulfide bridges are common in proteins that travel or operate in intracellular spaces, such as snake venoms and other toxins (e.g., conotoxins, charybdotoxin, bacterial

enterotoxins), peptide hormones, digestive enzymes, complement proteins, immunoglobulins, lysozymes, protease inhibitors (BPTI and its homologues, CMTI-III (Cucurbita maxima trypsin inhibitor III) and its homologues, hirudin, etc.) and milk proteins.

Disulfide bonds that close tight intrachain loops have been found in pepsin, thioredoxin, insulin A-chain, silk fibroin, and lipoamide dehydrogenase. The bridged cysteine residues are separated by one to four residues along the polypeptide chain. Model building, X-ray diffraction analysis, and NMR studies have shown that the α carbon path of such loops is usually flat and rigid.

There are two types of disulfide bridges in immunoglobulins. One is the conserved intrachain bridge, spanning about 60 to 70 amino acid residues and found, repeatedly, in almost every immunoglobulin domain. Buried deep between the opposing β sheets, these bridges are shielded from solvent and ordinarily can be reduced only in the presence of denaturing agents. The remaining disulfide bridges are mainly interchain bonds and are located on the surface of the molecule; they are accessible to solvent and relatively easily reduced (STEI85). The disulfide bridges of the micro-proteins of the present invention are intrachain linkages between cysteines having much smaller chain spacings.

When a micro-protein contains a plurality of disulfide bonds, it is preferable that at least two cysteines be clustered, i.e., are immediately adjacent along the chain (-C-C-) or are separated by a single amino acid (-C-X-C-). In either case, the two clustered cysteines become unable to pair with each other for steric reasons, and the number of realizable topologies is reduced.

An intrachain disulfide bridge connecting amino acids 3 and 8 of a 16 residue polypeptide will be said herein to have a span of 4. If amino acids 4 and 12 are also

disulfide bonded, then they form a second span of 7. Together, the four cysteines divide the polypeptide into four intercysteine segments (1-2, 5-7, 9-11, and 13-16). (Note that there is no segment between Cys3 and Cys4.) The connectivity pattern of a crosslinked micro-protein is a simple description of the relative location of the termini of the crosslinks. For example, for a micro-protein with two disulfide bonds, the connectivity pattern "1-3, 2-4" means that the first crosslinked cysteine is disulfide bonded to the third crosslinked cysteine (in the primary sequence), and the second to the fourth.

The degree to which the crosslink constrains the conformational freedom of the mini-protein, and the degree to which it stabilizes the mini-protein, may be assessed by a number of means. These include absorption spectroscopy (which can reveal whether an amino acid is buried or exposed), circular dichroism studies (which provides a general picture of the helical content of the protein), nuclear magnetic resonance imaging (which reveals the number of nuclei in a particular chemical environment as well as the mobility of nuclei), and X-ray or neutron diffraction analysis of protein crystals. The stability of the mini-protein may be ascertained by monitoring the changes in absorption at various wavelengths as a function of temperature, pH, etc.; buried residues become exposed as the protein unfolds. Similarly, the unfolding of the mini-protein as a result of denaturing conditions results in changes in NMR line positions and widths. Circular dichroism (CD) spectra are extremely sensitive to conformation.

The variegated disulfide-bonded micro-proteins of the present invention fall into several classes.

Class I micro-proteins are those featuring a single pair of cysteines capable of interacting to form a disulfide bond, said bond having a span of no more than about nine

residues. This disulfide bridge preferably has a span of at least two residues; this is a function of the geometry of the disulfide bond. When the spacing is two or three residues, one residue is preferably glycine in order to reduce the strain on the bridged residues. The upper limit on spacing is less precise, however, in general, the greater the spacing, the less the constraint on conformation imposed on the linearly intermediate amino acid residues by the disulfide bond.

The main chain of such a peptide has very little freedom, but is not stressed. The free energy released when the disulfide forms exceeds the free energy lost by the main-chain when locked into a conformation that brings the cysteines together. Having lost the free energy of disulfide formation, the proximal ends of the side groups are held in more or less fixed relation to each other. When binding to a target, the domain does not need to expend free energy getting into the correct conformation. The domain can not jump into some other conformation and bind a non-target.

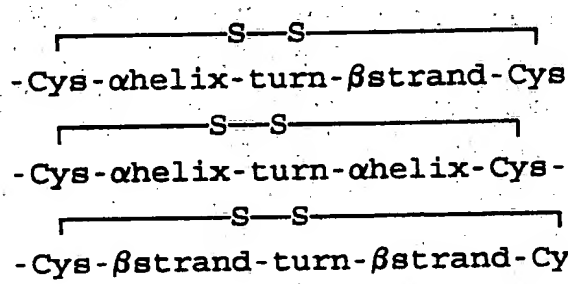
A disulfide bridge with a span of 4 or 5 is especially preferred. If the span is increased to 6, the constraining influence is reduced. In this case, we prefer that at least one of the enclosed residues be an amino acid that imposes restrictions on the main-chain geometry. Proline imposes the most restriction. Valine and isoleucine restrict the main chain to a lesser extent. The preferred position for this constraining non-cysteine residue is adjacent to one of the invariant cysteines, however, it may be one of the other bridged residues. If the span is seven, we prefer to include two amino acids that limit main-chain conformation. These amino acids could be at any of the seven positions, but are preferably the two bridged residues that are immediately adjacent to the cysteines. If the span is eight

or nine, additional constraining amino acids may be provided.

While a class I micro-protein may have up to 40 amino acids, more preferably it is no more than 20 amino acids.

The disulfide bond of a class I micro-proteins is exposed to solvent. Thus, one usually should avoid exposing the variegated population of GPs that display class I micro-proteins to reagents that rupture disulfides.

Class II micro-proteins are those featuring a single disulfide bond having a span of greater than nine amino acids. The bridged amino acids form secondary structures which help to stabilize their conformation. Preferably, these intermediate amino acids form hairpin supersecondary structures such as those schematized below:



Based on studies of known proteins, one may calculate the propensity of a particular residue, or of a particular dipeptide or tripeptide, to be found in an α helix, β strand or reverse turn. The normalized frequencies of occurrence of the amino acid residues in these secondary structures is given in Table 6-4 of CREI84. For a more detailed treatment on the prediction of secondary structure from the amino acid sequence, see Chapter 6 of SCHU79.

In designing a suitable hairpin structure, one may copy an actual structure from a protein whose three-dimensional conformation is known, design the structure using frequency data, or combine the two approaches. Preferably, one or more actual structures are used as a model, and the

frequency data is used to determine which mutations can be made without disrupting the structure.

Preferably, no more than three amino acids lie between the cysteine and the beginning or end of the α helix or β strand.

More complex structures (such as a double hairpin) are also possible.

Class IIIa micro-proteins are those featuring two disulfide bonds. They optionally may also feature secondary structures such as those discussed above with regard to Class II micro-proteins. With two disulfide bonds, there are three possible topologies; if desired, the number of realizable disulfide bonding topologies may be reduced by clustering cysteines as in heat-stable enterotoxin ST-Ia.

Class IIIb micro-proteins are those featuring three or more disulfide bonds and preferably at least one cluster of cysteines as previously described.

Metal Finger Mini-Proteins. The present invention also relates to mini-proteins which are not crosslinked by disulfide bonds, e.g., analogues of finger proteins. Finger proteins are characterized by finger structures in which a metal ion is coordinated by two Cys and two His residues, forming a tetrahedral arrangement around it. The metal ion is most often zinc(II), but may be iron, copper, cobalt, etc. The "finger" has the consensus sequence (Phe or Tyr)-(1 AA)-Cys-(2-4 AAs)-Cys-(3 AAs)-Phe-(5 AAs)-Leu-(2 AAs)-His-(3 AAs)-His-(5 AAs)(BERG88; GIBS88). While finger proteins typically contain many repeats of the finger motif, it is known that a single finger will fold in the presence of zinc ions (FRAN87; PARR88). There is some dispute as to whether two fingers are necessary for binding to DNA. The present invention encompasses mini-proteins with either one or two fingers. Other combinations of side groups can lead to formation of crosslinks involving multivalent metal ions. Summers (SUMM91), for example, reports an 18-amino-acid mini

protein found in the capsid protein of HIV-1-F1 and having three cysteines and one histidine that bind a zinc atom. It is to be understood that the target need not be a nucleic acid.

G. Modified PBSs

There exist a number of enzymes and chemical reagents that can selectively modify certain side groups of proteins, including: a) protein-tyrosine kinase, Ellmans reagent, methyl transferases (that methylate GLU side groups), serine kinases, proline hydroxyases, vitamin-K dependent enzymes that convert GLU to GLA, maleic anhydride, and alkylating agents. Treatment of the variegated population of GP(PBD)s with one of these enzymes or reagents will modify the side groups affected by the chosen enzyme or reagent. Enzymes and reagents that do not kill the GP are much preferred. Such modification of side groups can directly affect the binding properties of the displayed PBDs. Using affinity separation methods, we enrich for the modified GPs that bind the predetermined target. Since the active binding domain is not entirely genetically specified, we must repeat the post-morphogenesis modification at each enrichment round. This approach is particularly appropriate with mini-protein IPBDs because we envision chemical synthesis of these SBDs.

III. VARIEGATION STRATEGY -- MUTAGENESIS TO OBTAIN POTENTIAL BINDING DOMAINS WITH DESIRED DIVERSITY

III.A. Generally

When the number of different amino acid sequences obtainable by mutation of the domain is large when compared to the number of different domains which are displayable in detectable amounts, the efficiency of the forced evolution is greatly enhanced by careful choice of which residues are to be varied. First, residues of a known protein which are likely to affect its binding activity (e.g., surface residues) and not likely to unduly degrade its stability are

identified. Then all or some of the codons encoding these residues are varied simultaneously to produce a variegated population of DNA. Groups of surface residues that are close enough together on the surface to touch one molecule of target simultaneously are preferred sets for simultaneous variegation. The variegated population of DNA is used to express a variety of potential binding domains, whose ability to bind the target of interest may then be evaluated.

The method of the present invention is thus further distinguished from other methods in the nature of the highly variegated population that is produced and from which novel binding proteins are selected. We force the displayed potential binding domain to sample the nearby "sequence space" of related amino-acid sequences in an efficient, organized manner. Four goals guide the various variegation plans used herein, preferably: 1) a very large number (e.g. 10^7) of variants is available, 2) a very high percentage of the possible variants actually appears in detectable amounts, 3) the frequency of appearance of the desired variants is relatively uniform, and 4) variation occurs only at a limited number of amino-acid residues, most preferably at residues having side groups directed toward a common region on the surface of the potential binding domain.

This is to be distinguished from the simple use of indiscriminate mutagenic agents such as radiation and hydroxylamine to modify a gene, where there is no (or very oblique) control over the site of mutation. Many of the mutations will affect residues that are not a part of the binding domain. When chemical mutagens are directed toward the whole genome, most mutations occur in genes other than the one encoding the potential binding domain. Moreover, since at a reasonable level of mutagenesis, any modified codon is likely to be characterized by a single base change,

only a limited and biased range of possibilities will be explored. Equally remote is the use of site-specific mutagenesis techniques employing mutagenic oligonucleotides of nonrandomized sequence, since these techniques do not lend themselves to the production and testing of a large number of variants. While focused random mutagenesis techniques are known, the importance of controlling the distribution of variation has been largely overlooked.

The potential binding domains are first designed at the amino acid level. Once we have identified which residues are to be mutagenized, and which mutations to allow at those positions, we may then design the variegated DNA which is to encode the various PBDs so as to assure that there is a reasonable probability that if a PBD has an affinity for the target, it will be detected. Of course, the number of independent transformants obtained and the sensitivity of the affinity separation technology will impose limits on the extent of variegation possible within any single round of variegation.

There are many ways to generate diversity in a protein. (See RICH86, CARU85, and OLIP86.) At one extreme, we vary a few residues of the protein as much as possible (inter alia see CARU85, CARU87, RICH86, and WHAR86). We will call this approach "Focused Mutagenesis". A typical "Focused Mutagenesis" strategy is to pick a set of five to seven residues and vary each through 13-20 possibilities. An alternative plan of mutagenesis ("Diffuse Mutagenesis") is to vary many more residues through a more limited set of choices (See VERS86a and PAKU86). The variegation pattern adopted may fall between these extremes, e.g., two residues varied through all twenty amino acids, two more through only two possibilities, and a fifth into ten of the twenty amino acids.

There is no fixed limit on the number of codons which can be mutated simultaneously. However, it is desirable to

adopt a mutagenesis strategy which results in a reasonable probability that a possible PBD sequence is in fact displayed by at least one genetic package. Preferably, the probability that a mutein encoded by the vgDNA and composed of the least favored amino acids at each variegated position will be displayed by at least one independent transformant in the library is at least 0.50, and more preferably at least 0.90. (Muteins composed of more favored amino acids would of course be more likely to occur in the same library.)

Preferably, the variegation is such as will cause a typical transformant population to display 10^6-10^7 different amino acid sequences by means of preferably not more than 10-fold more (more preferably not more than 3-fold) different DNA sequences.

For a Class I micro-protein that lacks α helices and β strands, one will, in any given round of mutation, preferably variegate each of 4-8 non-cysteine codons so that they each encode at least eight of the 20 possible amino acids. The variegation at each codon could be customized to that position. Preferably, cysteine is not one of the potential substitutions, though it is not excluded.

When the mini-protein is a metal finger protein, in a typical variegation strategy, the two Cys and two His residues, and optionally also the aforementioned Phe/Tyr, Phe and Leu residues, are held invariant and a plurality (usually 5-10) of the other residues are varied.

When the micro-protein is of the type featuring one or more α helices and β strands, the set of potential amino acid modifications at any given position is picked to favor those which are less likely to disrupt the secondary structure at that position. Since the number of possibilities at each variable amino acid is more limited, the total

number of variable amino acids may be greater without altering the sampling efficiency of the selection process.

For class III micro-proteins, preferably not more than 20 and more preferably 5-10 codons will be variegated. However, if diffuse mutagenesis is employed, the number of codons which are variegated can be higher.

While variegation normally will involve the substitution of one amino acid for another at a designated variable codon, it may involve the insertion or deletion of amino acids as well.

III.B. Identification of Residues to be Varied

We now consider the principles that guide our choice of residues of the IPBD to vary. A key concept is that only structured proteins exhibit specific binding, i.e. can bind to a particular chemical entity to the exclusion of most others. Thus the residues to be varied are chosen with an eye to preserving the underlying IPBD structure. Substitutions that prevent the PBD from folding will cause GPs carrying those genes to bind indiscriminately so that they can easily be removed from the population. Substitutions of amino acids that are exposed to solvent are less likely to affect the 3D structure than are substitutions at internal loci. (See PAKU86, REID88a, EISE85, SCHU79, p169-171 and CREI84, p239-245, 314-315). Internal residues are frequently conserved and the amino acid type cannot be changed to a significantly different type without substantial risk that the protein structure will be disrupted. Nevertheless, some conservative changes of internal residues, such as I to L or F to Y, are tolerated. Such conservative changes subtly affect the placement and dynamics of adjacent protein residues and such "fine tuning" may be useful once an SBD is found. Insertions and deletions are more readily tolerated in loops than elsewhere. (THOR88).

Data about the IPBD and the target that are useful in deciding which residues to vary in the variegation cycle include: 1) 3D structure, or at least a list of residues on the surface of the IPBD, 2) list of sequences homologous to IPBD, and 3) model of the target molecule or a stand-in for the target.

III.C. Determining the Substitution Set for Each Parental Residue

Having picked which residues to vary, we now decide the range of amino acids to allow at each variable residue. The total level of variegation is the product of the number of variants at each varied residue. Each varied residue can have a different scheme of variegation, producing 2 to 20 different possibilities. The set of amino acids which are potentially encoded by a given variegated codon are called its "substitution set".

The computer that controls a DNA synthesizer, such as the Milligen 7500, can be programmed to synthesize any base of an oligo-nt with any distribution of nts by taking some nt substrates (e.g. nt phosphoramidites) from each of two or more reservoirs. Alternatively, nt substrates can be mixed in any ratios and placed in one of the extra reservoir for so called "dirty bottle" synthesis. Each codon could be programmed differently. The "mix" of bases at each nucleotide position of the codon determines the relative frequency of occurrence of the different amino acids encoded by that codon.

Simply variegated codons are those in which those nucleotide positions which are degenerate are obtained from a mixture of two or more bases mixed in equimolar proportions. These mixtures are described in this specification by means of the standardized "ambiguous nucleotide" code. In this code, for example, in the degenerate codon "SNT", "S" denotes an equimolar mixture of bases G and C, "N", an

equimolar mixture of all four bases, and "T", the single invariant base thymidine.

Complexly variegated codons are those in which at least one of the three positions is filled by a base from an other than equimolar mixture of two of more bases.

Either simply or complexly variegated codons may be used to achieve the desired substitution set.

If we have no information indicating that a particular amino acid or class of amino acid is appropriate, we strive to substitute all amino acids with equal probability because representation of one mini-protein above the detectable level is wasteful. Equal amounts of all four nts at each position in a codon (NNN) yields the amino acid distribution in which each amino acid is present in proportion to the number of codons that code for it. This distribution has the disadvantage of giving two basic residues for every acidic residue. In addition, six times as much R, S, and L as W or M occur. If five codons are synthesized with this distribution, each of the 243 sequences encoding some combination of L, R, and S are 7776-times more abundant than each of the 32 sequences encoding some combination of W and M. To have five Ws present at detectable levels, we must have each of the (L,R,S) sequences present in 7776-fold excess.

Particular amino acid residues can influence the tertiary structure of a defined polypeptide in several ways, including by:

a) affecting the flexibility of the polypeptide main chain,
b) adding hydrophobic groups,
c) adding charged groups,
d) allowing hydrogen bonds, and
e) forming cross-links, such as disulfides, chelation to metal ions, or bonding to prosthetic groups.

Lundeen (LUND86) has tabulated the frequencies of amino acids in helices, β strands, turns, and coil in proteins of known 3D structure and has distinguished between CYSs having free thiol groups and half cystines. He reports that free CYS is found most often in helixes while half cystines are found more often in β sheets. Half cystines are, however, regularly found in helices. Pease et al. (PEAS90) constructed a peptide having two cystines; one end of each is in a very stable α helix. Apamin has a similar structure (WEMM83, PEAS88).

Flexibility:

GLY is the smallest amino acid, having two hydrogens attached to the C_α. Because GLY has no C_β, it confers the most flexibility on the main chain. Thus GLY occurs very frequently in reverse turns, particularly in conjunction with PRO, ASP, ASN, SER, and THR.

The amino acids ALA, SER, CYS, ASP, ASN, LEU, MET, PHE, TYR, TRP, ARG, HIS, GLU, GLN, and LYS have unbranched β carbons. Of these, the side groups of SER, ASP, and ASN frequently make hydrogen bonds to the main chain and so can take on main-chain conformations that are energetically unfavorable for the others. VAL, ILE, and THR have branched β carbons which makes the extended main-chain conformation more favorable. Thus VAL and ILE are most often seen in β sheets. Because the side group of THR can easily form hydrogen bonds to the main chain, it has less tendency to exist in a β sheet.

The main chain of proline is particularly constrained by the cyclic side group. The ϕ angle is always close to -60°. Most prolines are found near the surface of the protein.

Charge:

LYS and ARG carry a single positive charge at any pH below 10.4 or 12.0, respectively. Nevertheless, the

methylene groups, four and three respectively, of these amino acids are capable of hydrophobic interactions. The guanidinium group of ARG is capable of donating five hydrogens simultaneously, while the amino group of LYS can donate only three. Furthermore, the geometries of these groups is quite different, so that these groups are often not interchangeable.

ASP and GLU carry a single negative charge at any pH above ~4.5 and 4.6, respectively. Because ASP has but one methylene group, few hydrophobic interactions are possible. The geometry of ASP lends itself to forming hydrogen bonds to main-chain nitrogens which is consistent with ASP being found very often in reverse turns and at the beginning of helices. GLU is more often found in α helices and particularly in the amino-terminal portion of these helices because the negative charge of the side group has a stabilizing interaction with the helix dipole (NICH88, SALI88).

HIS has an ionization pK in the physiological range, viz. 6.2. This pK can be altered by the proximity of charged groups or of hydrogen donators or acceptors. HIS is capable of forming bonds to metal ions such as zinc, copper, and iron.

Hydrogen bonds:

Aside from the charged amino acids, SER, THR, ASN, GLN, TYR, and TRP can participate in hydrogen bonds.

Cross links:

The most important form of cross link is the disulfide bond formed between the thiols of CYS residues. In a suitably oxidizing environment, these bonds form spontaneously. These bonds can greatly stabilize a particular conformation of a protein or mini-protein. When a mixture of oxidized and reduced thiol reagents are present, exchange reactions take place that allow the most stable conformation to predominate. Concerning disulfides

in proteins and peptides, see also KATZ90, MATS89, PERR84, PERR86, SAUE86, WELL86, JANA89, HORV89, KISH85, and SCHN86.

Other cross links that form without need of specific enzymes include:

1)	$(CYS)_4$:Fe	Rubredoxin (in CREI84, P.376)
2)	$(CYS)_4$:Zn	Aspartate Transcarbamylase (in CREI84, P.376) and Zn-fingers (HARD90)
3)	$(HIS)_2(MET)(CYS)$:Cu	Azurin (in CREI84, P.376) and Basic "Blue" Cu Cucumber protein (GUSS88)
4)	$(HIS)_4$:Cu	CuZn superoxide dismutase
5)	$(CYS)_4$:(Fe_4S_4)	Ferredoxin (in CREI84, P.376)
6)	$(CYS)_2(HIS)_2$:Zn	Zinc-fingers (GIBS88, SUMM91)
7)	$(CYS)_3(HIS)$:Zn	Zinc-fingers (GAUS87, GIBS88)

Cross links having $(HIS)_2(MET)(CYS)$:Cu has the potential advantage that HIS and MET can not form other cross links without Cu.

Simply Variegated Codons

The following simply variegated codons are useful because they encode a relatively balanced set of amino acids:

1) SNT which encodes the set [L,P,H,R,V,A,D,G]: a) one acidic (D) and one basic (R), b) both aliphatic (L,V) and aromatic hydrophobics (H), c) large (L,R,H) and small (G,A) side groups, d) rigid (P) and flexible (G) amino acids, e) each amino acid encoded once.
2) RNG which encodes the set [M,T,K,R,V,A,E,G]: a) one acidic and two basic (not optimal, but acceptable), b) hydrophilics and hydrophobics, c) each amino acid encoded once.
3) RMG which encodes the set [T,K,A,E]: a) one acidic, one basic, one neutral hydrophilic, b) three favor α helices, c) each amino acid encoded once.

4) VNT which encodes the set [L,P,H,R,I,T,N,S,V,A,D,G]: a) one acidic, one basic, b) all classes: charged, neutral hydrophilic, hydrophobic, rigid and flexible, etc., c) each amino acid encoded once.

5) RRS which encodes the set [N,S,K,R,D,E,G²]: a) two acidics, two basics, b) two neutral hydrophilics, c) only glycine encoded twice.

6) NNT which encodes the set [F,S,Y,C,L,P,H,R,I,T,N,V,A-,D,G]: a) sixteen DNA sequences provide fifteen different amino acids; only serine is repeated, all others are present in equal amounts (This allows very efficient sampling of the library.), b) there are equal numbers of acidic and basic amino acids (D and R, once each), c) all major classes of amino acids are present: acidic, basic, aliphatic hydrophobic, aromatic hydrophobic, and neutral hydrophilic.

7) NNG, which encodes the set [L²,R²,S,W,P,Q,M,T,K,V,A,-E,G, stop]: a) fair preponderance of residues that favor formation of α-helices [L,M,A,Q,K,E; and, to a lesser extent, S,R,T]; b) encodes 13 different amino acids. (VHG encodes a subset of the set encoded by NNG which encodes 9 amino acids in nine different DNA sequences, with equal acids and bases, and 5/9 being α helix-favoring.)

For the initial variegation, NNT is preferred, in most cases. However, when the codon is encoding an amino acid to be incorporated into an α helix, NNG is preferred.

Below, we analyze several simple variegations as to the efficiency with which the libraries can be sampled.

Libraries of random hexapeptides encoded by $(NNK)^6$ have been reported (SCOT90, CWIR90). Table 130 shows the expected behavior of such libraries. NNK produces single codons for PHE, TYR, CYS, TRP, HIS, GLN, ILE, MET, ASN, LYS, ASP, and GLU (α set); two codons for each of VAL, ALA, PRO,

THR, and GLY (Φ set); and three codons for each of LEU, ARG, and SER (Ω set). We have separated the 64,000,000 possible sequences into 28 classes, shown in Table 130A, based on the number of amino acids from each of these sets. The largest class is ΦΩαααα with ~14.6% of the possible sequences. Aside from any selection, all the sequences in one class have the same probability of being produced. Table 130B shows the probability that a given DNA sequence taken from the $(NNK)^6$ library will encode a hexapeptide belonging to one of the defined classes; note that only ~6.3% of DNA sequences belong to the ΦΩαααα class.

Table 130C shows the expected numbers of sequences in each class for libraries containing various numbers of independent transformants (viz. 10^6, $3\cdot10^6$, 10^7, $3\cdot10^7$, 10^8, $3\cdot10^8$, 10^9, and $3\cdot10^9$). At 10^6 independent transformants (ITs), we expect to see 56% of the ΩΩΩΩΩΩ class, but only 0.1% of the αααααα class. The vast majority of sequences seen come from classes for which less than 10% of the class is sampled. Suppose a peptide from, for example, class ΦΦΩΩαα is isolated by fractionating the library for binding to a target. Consider how much we know about peptides that are related to the isolated sequence. Because only 4% of the ΦΦΩΩαα class was sampled, we can not conclude that the amino acids from the Ω set are in fact the best from the Ω set. We might have LEU at position 2, but ARG or SER could be better. Even if we isolate a peptide of the ΩΩΩΩΩΩ class, there is a noticeable chance that better members of the class were not present in the library.

With a library of 10^7 ITs, we see that several classes have been completely sampled, but that the αααααα class is only 1.1% sampled. At $7.6\cdot10^7$ ITs, we expect display of 50% of all amino-acid sequences, but the classes containing three or more amino acids of the α set are still poorly sampled. To achieve complete sampling of the $(NNK)^6$ library

requires about $3 \cdot 10^9$ ITs, 10-fold larger than the largest $(NNK)^6$ library so far reported.

Table 131 shows expectations for a library encoded by $(NNT)^4(NNG)^2$. The expectations of abundance are independent of the order of the codons or of interspersed unvaried codons. This library encodes 0.133 times as many amino-acid sequences, but there are only 0.0165 times as many DNA sequences. Thus $5.0 \cdot 10^7$ ITs (i.e. 60-fold fewer than required for $(NNK)^6$) gives almost complete sampling of the library. The results would be slightly better for $(NNT)^6$ and slightly, but not much, worse for $(NNG)^6$. The controlling factor is the ratio of DNA sequences to amino-acid sequences.

Table 132 shows the ratio of #DNA sequences/#AA sequences for codons NNK, NNT, and NNG. For NNK and NNG, we have assumed that the PBD is displayed as part of an essential gene, such as gene III in Ff phage, as is indicated by the phrase "assuming stops vanish". It is not in any way required that such an essential gene be used. If a non-essential gene is used, the analysis would be slightly different; sampling of NNK and NNG would be slightly less efficient. Note that $(NNT)^6$ gives 3.6-fold more amino-acid sequences than $(NNK)^5$ but requires 1.7-fold fewer DNA sequences. Note also that $(NNT)^7$ gives twice as many amino-acid sequences as $(NNK)^6$, but 3.3-fold fewer DNA sequences.

Thus, while it is possible to use a simple mixture (NNS, NNK or NNN) to obtain at a particular position all twenty amino acids, these simple mixtures lead to a highly biased set of encoded amino acids. This problem can be overcome by use of complexly variegated codons.

Complexly Variegated Codons

The nt distribution ("fxS") within the codon that allows all twenty amino acids and that yields the largest ratio of abundance of the least favored amino acid (1faa) to

that of the most favored amino acid (mfaa), subject to the constraints of equal abundances of acidic and basic amino acids, least possible number of stop codons, and, for convenience, the third base being T or G, is shown in Table 10A and yields DNA molecules encoding each type of amino acid with the abundances shown. Other complexly variegated codons are obtainable by relaxing one or more constraints.

Note that this chemistry encodes all twenty amino acids, with acidic and basic amino acids being equiprobable, and the most favored amino acid (serine) is encoded only 2.454 times as often as the least favored amino acid (tryptophan). The "fxS" vg codon improves sampling most for peptides containing several of the amino acids [F,Y,C,W,H-,Q,I,M,N,K,D,E] for which NNK or NNS provide only one codon. Its sampling advantages are most pronounced when the library is relatively small.

The results of omitting the requirements of equality of acids and bases and minimizing stop codons are shown in Table 10B.

The advantages of an NNT codon are discussed elsewhere in the present application. Unoptimized NNT provides 15 amino acids encoded by only 16 DNA sequences. It is possible to improve on NNT with the distribution shown in Table 10C, which gives five amino acids (SER, LEU, HIS, VAL, ASP) in very nearly equal amounts. A further eight amino acids (PHE, TYR, ILE, ASN, PRO, ALA, ARG, GLY) are present at 78% the abundance of SER. THR and CYS remain at half the abundance of SER. When variegating DNA for disulfide-bonded micro-proteins, it is often desirable to reduce the prevalence of CYS. This distribution allows 13 amino acids to be seen at high level and gives no stops; the optimized fxS distribution allows only 11 amino acids at high prevalence.

The NNG codon can also be optimized. Table 10D shows an approximately optimized ([ALA] ~ [ARG]) NNG codon. There

are, under this variegation, four equally most favored amino acids: LEU, ARG, ALA, and GLU. Note that there is one acidic and one basic amino acid in this set. There are two equally least favored amino acids: TRP and MET. The ratio of lfaa/mfaa is 0.5258. If this codon is repeated six times, peptides composed entirely of TRP and MET are 2% as common as peptides composed entirely of the most favored amino acids. We refer to this as "the prevalence of $(TRP/MET)^6$ in optimized NNG^6 vgDNA".

When synthesizing vgDNA by the "dirty bottle" method, it is sometimes desirable to use only a limited number of mixes. One very useful mixture is called the "optimized NNS mixture" in which we average the first two positions of the fxS mixture: $T_1 = 0.24$, $C_1 = 0.17$, $A_1 = 0.33$, $G_1 = 0.26$, the second position is identical to the first, $C_3 = G_3 = 0.5$. This distribution provides the amino acids ARG, SER, LEU, GLY, VAL, THR, ASN, and LYS at greater than 5% plus ALA, ASP, GLU, ILE, MET, and TYR at greater than 4%.

An additional complexly variegated codon is of interest. This codon is identical to the optimized NNT codon at the first two positions and has T:G::90:10 at the third position. This codon provides thirteen amino acids (ALA, ILE, ARG, SER, ASP, LEU, VAL, PHE, ASN, GLY, PRO, TYR, and HIS) at more than 5.5%. THR at 4.3% and CYS at 3.9% are more common than the LFAAs of NNK (3.125%). The remaining five amino acids are present at less than 1%. This codon has the feature that all amino acids are present; sequences having more than two of the low-abundance amino acids are rare. When we isolate an SBD using this codon, we can be reasonably sure that the first 13 amino acids were tested at each position. A similar codon, based on optimized NNG, could be used.

Table 10E shows some properties of an unoptimized NNS (or NNK) codon. Note that there are three equally most-

favored amino acids: ARG, LEU, and SER. There are also twelve equally least favored amino acids: PHE, ILE, MET, TYR, HIS, GLN, ASN, LYS, ASP, GLU, CYS, and TRP. Five amino acids (PRO, THR, ALA, VAL, GLY) fall in between. Note that a six-fold repetition of NNS gives sequences composed of the amino acids [PHE, ILE, MET, TYR, HIS, GLN, ASN, LYS, ASP, GLU, CYS, and TRP] at only ~0.1% of the sequences composed of [ARG, LEU, and SER]. Not only is this ~20-fold lower than the prevalence of $(TRP/MET)^6$ in optimized NNG^6 vgDNA, but this low prevalence applies to twelve amino acids.

Diffuse Mutagenesis

Diffuse Mutagenesis can be applied to any part of the protein at any time, but is most appropriate when some binding to the target has been established. Diffuse Mutagenesis can be accomplished by spiking each of the pure nts activated for DNA synthesis (e.g. nt-phosphoramidites) with a small amount of one or more of the other activated nts. Preferably, the level of spiking is set so that only a small percentage (1% to .00001%, for example) of the final product will contain the initial DNA sequence. This will insure that many single, double, triple, and higher mutations occur, but that recovery of the basic sequence will be a possible outcome.

III.D. Special Considerations Relating to Variegation of Micro-Proteins with Essential Cysteines

Several of the preferred simple or complex variegated codons encode a set of amino acids which includes cysteine. This means that some of the encoded binding domains will feature one or more cysteines in addition to the invariant disulfide-bonded cysteines. For example, at each NNT-encoded position, there is a one in sixteen chance of obtaining cysteine. If six codons are so varied, the fraction of domains containing additional cysteines is 0.33. Odd numbers of cysteines can lead to complications, see

Perry and Wetzel (PERR84). On the other hand, many disulfide-containing proteins contain cysteines that do not form disulfides, e.g. trypsin. The possibility of unpaired cysteines can be dealt with in several ways:

First, the variegated phage population can be passed over an immobilized reagent that strongly binds free thiols, such as SulfoLink (catalogue number 44895 H from Pierce Chemical Company, Rockford, Illinois, 61105). Another product from Pierce is TNB-Thiol Agarose (Catalogue Code 20409 H). BioRad sells Affi-Gel 401 (catalogue 153-4599) for this purpose.

Second, one can use a variegation that excludes cysteines, such as:

NHT that gives [F,S,Y,L,P,H,I,T,N,V,A,D],

VNS that gives [L^2,P^2,H,Q,R^3,I,M,T^2,N,K,S,V^2,A^2,E,D,G^2],

NNG that gives [L^2,S,W,P,Q,R^2,M,T,K,R,V,A,E,G,stop],

SNT that gives [L,P,H,R,V,A,D,G],

RNG that gives [M,T,K,R,V,A,E,G],

RMG that gives [T,K,A,E],

VNT that gives [L,P,H,R,I,T,N,S,V,A,D,G], or

RRS that gives [N,S,K,R,D,E,G^2].

However, each of these schemes has one or more of the disadvantages, relative to NNT: a) fewer amino acids are allowed, b) amino acids are not evenly provided, c) acidic and basic amino acids are not equally likely), or d) stop codons occur. Nonetheless, NNG, NHT, and VNT are almost as useful as NNT. NNG encodes 13 different amino acids and one stop signal. Only two amino acids appear twice in the 16-fold mix.

Thirdly, one can enrich the population for binding to the preselected target, and evaluate selected sequences post hoc for extra cysteines. Those that contain more cysteines than the cysteines provided for conformational constraint

may be perfectly usable. It is possible that a disulfide linkage other than the designed one will occur. This does not mean that the binding domain defined by the isolated DNA sequence is in any way unsuitable. The suitability of the isolated domains is best determined by chemical and biochemical evaluation of chemically synthesized peptides.

Lastly, one can block free thiols with reagents, such as Ellman's reagent, iodoacetate, or methyl iodide, that specifically bind free thiols and that do not react with disulfides, and then leave the modified phage in the population. It is to be understood that the blocking agent may alter the binding properties of the micro-protein; thus, one might use a variety of blocking reagent in expectation that different binding domains will be found. The variegated population of thiol-blocked genetic packages are fractionated for binding. If the DNA sequence of the isolated binding micro-protein contains an odd number of cysteines, then synthetic means are used to prepare micro-proteins having each possible linkage and in which the odd thiol is appropriately blocked. Nishiuchi (NISH82, NISH86, and works cited therein) disclose methods of synthesizing peptides that contain a plurality of cysteines so that each thiol is protected with a different type of blocking group. These groups can be selectively removed so that the disulfide pairing can be controlled. We envision using such a scheme with the alteration that one thiol either remains blocked, or is unblocked and then reblocked with a different reagent.

III.E. Planning the Second and Later Rounds of Variegation

The method of the present invention allows efficient accumulation of information concerning the amino-acid sequence of a binding domain having high affinity for a predetermined target. Although one may obtain a highly useful binding domain from a single round of variegation and affinity enrichment, we expect that multiple rounds will be

needed to achieve the highest possible affinity and specificity.

If the first round of variegation results in some binding to the target, but the affinity for the target is still too low, further improvement may be achieved by variegation of the SBDs. Preferably, the process is progressive, i.e. each variegation cycle produces a better starting point for the next variegation cycle than the previous cycle produced. Setting the level of variegation such that the ppbd and many sequences related to the ppbd sequence are present in detectable amounts ensures that the process is progressive.

If the level of variegation is so high that the ppbd sequence is present at such low levels that there is an appreciable chance that no transformant will display the PPBD, then the best SBD of the next round could be worse than the PPBD. At excessively high level of variegation, each round of mutagenesis is independent of previous rounds and there is no assurance of progressivity. This approach can lead to valuable binding proteins, but repetition of experiments with this level of variegation will not yield progressive results. Excessive variation is not preferred.

Progressivity is not an all-or-nothing property. So long as most of the information obtained from previous variegation cycles is retained and many different surfaces that are related to the PPBD surface are produced, the process is progressive.

If the level of variegation in the previous variegation cycle was correctly chosen, then the amino acids selected to be in the residues just varied are the ones best determined. The environment of other residues has changed, so that it is appropriate to vary them again. Because there are often more residues of interest than can be varied simultaneously, we may continue by picking residues that either have never

been varied (highest priority) or that have not been varied for one or more cycles.

Use of NNT or NNG variegated codons leads to very efficient sampling of variegated libraries because the ratio of (different amino-acid sequences)/(different DNA sequences) is much closer to unity than it is for NNK or even the optimized vg codon (fxS). Nevertheless, a few amino acids are omitted in each case. Both NNT and NNG allow members of all important classes of amino acids: hydrophobic, hydrophilic, acidic, basic, neutral hydrophilic, small, and large. After selecting a binding domain, a subsequent variegation and selection may be desirable to achieve a higher affinity or specificity. During this second variegation, amino acid possibilities overlooked by the preceding variegation may be investigated.

A few examples may be helpful. Suppose we obtained PRO using NNT. This amino acid is available with either NNT or NNG. We can be reasonably sure that PRO is the best amino acid from the set [PRO, LEU, VAL, THR, ALA, ARG, GLY, PHE, TYR, CYS, HIS, ILE, ASN, ASP, SER]. We next might try a set that includes [PRO, TRP, GLN, MET, LYS, GLU]. The set allowed by NNG is the preferred set.

What if we obtained HIS instead? Histidine is aromatic and fairly hydrophobic and can form hydrogen bonds to and from the imidazole ring. Tryptophan is hydrophobic and aromatic and can donate a hydrogen to a suitable acceptor and was excluded by the NNT codon. Methionine was also excluded and is hydrophobic. Thus, one preferred course is to use the variegated codon HDS that allows [HIS, GLN, ASN, LYS, TYR, CYS, TRP, ARG, SER, GLY, <stop>].

If the first round of variegation is entirely unsuccessful, a different pattern of variegation should be used. For example, if more than one interaction set can be defined within a domain, the residues varied in the next round of variegation should be from a different set than

that probed in the initial variegation. If repeated failures are encountered, one may switch to a different IPBD.

IV. DISPLAY STRATEGY: DISPLAYING FOREIGN BINDING DOMAINS ON THE SURFACE OF A "GENETIC PACKAGE"

IV.A. General Requirements for Genetic Packages

In order to obtain the display of a multitude of different though related potential binding domains, applicants generate a heterogeneous population of replicable genetic packages each of which comprises a hybrid gene including a first DNA sequence which encodes a potential binding domain for the target of interest and a second DNA sequence which encodes a display means, such as an outer surface protein native to the genetic package but not natively associated with the potential binding domain (or the parental binding domain to which it is related) which causes the genetic package to display the corresponding chimeric protein (or a processed form thereof) on its outer surface.

The component of a population that exhibits the desired binding properties may be quite small, for example, one in 10^6 or less. Once this component of the population is separated from the non-binding components, it must be possible to amplify it. Culturing viable cells is the most powerful amplification of genetic material known and is preferred. Genetic messages can also be amplified in vitro, e.g. by PCR, but this is not the most preferred method.

Preferably, the GP can be: 1) genetically altered with reasonable facility to encode a potential binding domain, 2) maintained and amplified in culture, 3) manipulated to display the potential binding protein domain where it can interact with the target material during affinity separation, and 4) affinity separated while retaining the

genetic information encoding the displayed binding domain in recoverable form. Preferably, the GP remains viable after affinity separation. Preferred GPs are vegetative bacterial cells, bacterial spores and, especially, bacterial DNA viruses. Eukaryotic cells and eukaryotic viruses may be used as genetic packages, but are not preferred.

When the genetic package is a bacterial cell, or a phage which is assembled periplasmically, the display means has two components. The first component is a secretion signal which directs the initial expression product to the inner membrane of the cell (a host cell when the package is a phage). This secretion signal is cleaved off by a signal peptidase to yield a processed, mature, potential binding protein. The second component is an outer surface transport signal which directs the package to assemble the processed protein into its outer surface. Preferably, this outer surface transport signal is derived from a surface protein native to the genetic package.

For example, in a preferred embodiment, the hybrid gene comprises a DNA encoding a potential binding domain operably linked to a signal sequence (e.g., the signal sequences of the bacterial phoA or bla genes or the signal sequence of M13 phage geneIII) and to DNA encoding a coat protein (e.g., the M13 gene III or gene VIII proteins) of a filamentous phage (e.g., M13). The expression product is transported to the inner membrane (lipid bilayer) of the host cell, whereupon the signal peptide is cleaved off to leave a processed hybrid protein. The C-terminus of the coat protein-like component of this hybrid protein is trapped in the lipid bilayer, so that the hybrid protein does not escape into the periplasmic space. (This is typical of the wild-type coat protein.) As the single-stranded DNA of the nascent phage particle passes into the periplasmic space, it collects both wild-type coat protein and the hybrid protein from the lipid bilayer. The hybrid

protein is thus packaged into the surface sheath of the filamentous phage, leaving the potential binding domain exposed on its outer surface. (Thus, the filamentous phage, not the host bacterial cell, is the "replicable genetic package" in this embodiment.)

If a secretion signal is necessary for the display of the potential binding domain, in an especially preferred embodiment the bacterial cell in which the hybrid gene is expressed is of a "secretion-permissive" strain.

When the genetic package is a bacterial spore, or a phage (such as ΦX174 or λ) whose coat is assembled intracellularly, a secretion signal directing the expression product to the inner membrane of the host bacterial cell is unnecessary. In these cases, the display means is merely the outer surface transport signal, typically a derivative of a spore or phage coat protein.

Preferred OSPs for several GPs are given in Table 2. References to osp-ipbd fusions in this section should be taken to apply, mutatis mutandis, to osp-pbd and osp-sbd fusions as well.

IV.B. Phages for Use as GPs:

Periplasmically assembled phage are preferred when the IPBD is a disulfide-bonded micro-protein, as such IPBDs may not fold within a cell (these proteins may fold after the phage is released from the cell). Intracellularly assembled phage are preferred when the IPBD needs large or insoluble prosthetic groups (such as Fe_4S_4 clusters), since the IPBD may not fold if secreted because the prosthetic group is lacking in the periplasm.

When variegation is introduced, multiple infections could generate hybrid GPs that carry the gene for one PBD but have at least some copies of a different PBD on their surfaces; it is preferable to minimize this possibility by

infecting cells with phage under conditions resulting in a low multiple-of-infection (MOI).

Bacteriophages are excellent candidates for GPs because there is little or no enzymatic activity associated with intact mature phage, and because the genes are inactive outside a bacterial host, rendering the mature phage particles metabolically inert.

For a given bacteriophage, the preferred OSP is usually one that is present on the phage surface in the largest number of copies. Nevertheless, an OSP such as M13 gIII protein (5 copies/phage) may be an excellent choice as OSP to cause display of the PBD.

It is preferred that the wild-type osp gene be preserved. The ipbd gene fragment may be inserted either into a second copy of the recipient osp gene or into a novel engineered osp gene. It is preferred that the osp-ipbd gene be placed under control of a regulated promoter.

The user must choose a site in the candidate OSP gene for inserting a ipbd gene fragment. The coats of most bacteriophage are highly ordered. In such bacteriophage, it is important to retain in engineered OSP-IPBD fusion proteins those residues of the parental OSP that interact with other proteins in the virion. For M13 gVIII, we preferably retain the entire mature protein, while for M13 gIII, it might suffice to retain the last 100 residues (BASS90) (or even fewer). Such a truncated gIII protein would be expressed in parallel with the complete gIII protein, as gIII protein is required for phage infectivity.

The filamentous phage, which include M13, f1, fd, If1, Ike, Xf, Pf1, and Pf3, are of particular interest. The major coat protein is encoded by gene VIII. The 50 amino acid mature gene VIII coat protein is synthesized as a 73 amino acid precoat (ITOK79). The first 23 amino acids constitute a typical signal-sequence which causes the

nascent polypeptide to be inserted into the inner cell membrane.

An E. coli signal peptidase (SP-I) recognizes amino acids 18, 21, and 23, and, to a lesser extent, residue 22, and cuts between residues 23 and 24 of the precoat (KUHN85a, KUHN85b, OLIV87). After removal of the signal sequence, the amino terminus of the mature coat is located on the periplasmic side of the inner membrane; the carboxy terminus is on the cytoplasmic side. About 3000 copies of the mature 50 amino acid coat protein associate side-by-side in the inner membrane.

The sequence of gene VIII is known, and the amino acid sequence can be encoded on a synthetic gene, using lacUV5 promoter and used in conjunction with the $LacI^q$ repressor. The lacUV5 promoter is induced by IPTG. Mature gene VIII protein makes up the sheath around the circular ssDNA. The 3D structure of f1 virion is known at medium resolution; the amino terminus of gene VIII protein is on surface of the virion and is therefore a preferred atttachment site for the potential binding domain. A few modifications of gene VIII have been made and are discussed below. The 2D structure of M13 coat protein is implicit in the 3D structure. Mature M13 gene VIII protein has only one domain.

We have constructed a tripartite gene comprising:

1) DNA encoding a signal sequence directing secretion of parts (2) and (3) through the inner membrane,
2) DNA encoding the mature BPTI sequence, and
3) DNA encoding the mature M13 gVIII protein.

This gene causes BPTI to appear in active form on the surface of M13 phage.

The amino-acid sequence of M13 pre-coat (SCHA78), called AA_seq1, is

AA_seq1

```
              1    1    2   ||2   3    3    4    4    5
         5    0    5    0   \/5   0    5    0    5    0
MKKSLVLKASVAVATLVPMLSFAAEGDDPAKAAFNSLQASATEYIGYAWA

    5    6    6    7  7
    5    0    5    0  3
MVVVIVGATIGIKLFKKFTSKAS
```

The best site for inserting a novel protein domain into M13 CP is after A23 because SP-I cleaves the precoat protein after A23, as indicated by the arrow. Proteins that can be secreted will appear connected to mature M13 CP at its amino terminus. Because the amino terminus of mature M13 CP is located on the outer surface of the virion, the introduced domain will be displayed on the outside of the virion. The uncertainty of the mechanism by which M13CP appears in the lipid bilayer raises the possibility that direct insertion of <u>bpti</u> into gene <u>VIII</u> may not yield a functional fusion protein. It may be necessary to change the signal sequence of the fusion to, for example, the <u>phoA</u> signal sequence (MKQSTIALALLPLLFTPVTKA......) (MARK91). Marks <u>et al.</u> (MARK86) showed that the <u>phoA</u> signal peptide could direct mature BPTI to the <u>E. coli</u> periplasm.

Another vehicle for displaying the IPBD is by expressing it as a domain of a chimeric gene containing part or all of gene <u>III</u>. This gene encodes one of the minor coat proteins of M13. Genes VI, VII, and IX also encode minor coat proteins. Each of these minor proteins is present in about 5 copies per virion and is related to morphogenesis or infection. In contrast, the major coat protein is present in more than 2500 copies per virion. The gene VI, VII, and IX proteins are present at the ends of the virion; these three proteins are not post-translationally processed (RASC86).

The single-stranded circular phage DNA associates with about five copies of the gene III protein and is then extruded through the patch of membrane-associated coat protein in such a way that the DNA is encased in a helical sheath of protein (WEBS78). The DNA does not base pair (that would impose severe restrictions on the virus genome); rather the bases intercalate with each other independent of sequence.

Smith (SMIT85) and de la Cruz et al. (DELA88) have shown that insertions into gene III cause novel protein domains to appear on the virion outer surface. The mini-protein's gene may be fused to gene III at the site used by Smith and by de la Cruz et al., at a codon corresponding to another domain boundary or to a surface loop of the protein, or to the amino terminus of the mature protein.

All published works use a vector containing a single modified gene III of fd. Thus, all five copies of gIII are identically modified. Gene III is quite large (1272 b.p. or about 20% of the phage genome) and it is uncertain whether a duplicate of the whole gene can be stably inserted into the phage. Furthermore, all five copies of gIII protein are at one end of the virion. When bivalent target molecules (such as antibodies) bind a pentavalent phage, the resulting complex may be irreversible. Irreversible binding of the GP to the target greatly interferes with affinity enrichment of the GPs that carry the genetic sequences encoding the novel polypeptide having the highest affinity for the target.

To reduce the likelihood of formation of irreversible complexes, we may use a second, synthetic gene that encodes carboxy-terminal parts of III; the carboxy-terminal parts of the gene III protein cause it to assemble into the phage. For example, the final 29 residues (starting with the arginine specified by codon 398) may be enough to cause a fusion protein to assemble into the phage. Alternatively, one might include the final globular domain of mature gIII

protein, viz. the final 150 to 160 amino acids of gene III (BASS90). We might, for example, engineer a gene that consists of (from 5' to 3'):

1) a promoter (preferably regulated),
2) a ribosome-binding site,
3) an initiation codon,
4) a functional signal peptide directing secretion of parts (5) and (6) through the inner membrane,
5) DNA encoding an IPBD,
6) DNA encoding residues 275 through 424 of M13 gIII protein,
7) a translation stop codon, and
8) (optionally) a transcription stop signal.

We leave the wild-type gene III so that some unaltered gene III protein will be present. Alternatively, we may use gene VIII protein as the OSP and regulate the osp::ipbd fusion so that only one or a few copies of the fusion protein appear on the phage.

M13 gene VI, VII, and IX proteins are not processed after translation. The route by which these proteins are assembled into the phage have not been reported. These proteins are necessary for normal morphogenesis and infectivity of the phage. Whether these molecules (gene VI protein, gene VII protein, and gene IX protein) attach themselves to the phage: a) from the cytoplasm, b) from the periplasm, or c) from within the lipid bilayer, is not known. One could use any of these proteins to introduce an IPBD onto the phage surface by one of the constructions:

1) ipbd::pmcp,
2) pmcp::ipbd,
3) signal::ipbd::pmcp, and
4) signal::pmcp::ipbd.

where ipbd represents DNA coding on expression for the initial potential binding domain; pmcp represents DNA coding for one of the phage minor coat proteins, VI, VII, and IX;

signal represents a functional secretion signal peptide, such as the phoA signal (MKQSTIALALLPLLFTPVTKA); and "::" represents in-frame genetic fusion. The indicated fusions are placed downstream of a known promoter, preferably a regulated promoter such as lacUV5, tac, or trp. Fusions (1) and (2) are appropriate when the minor coat protein attaches to the phage from the cytoplasm or by autonomous insertion into the lipid bilayer. Fusion (1) is appropriate if the amino terminus of the minor coat protein is free and (2) is appropriate if the carboxy terminus is free. Fusions (3) and (4) are appropriate if the minor coat protein attaches to the phage from the periplasm or from within the lipid bilayer. Fusion (3) is appropriate if the amino terminus of the minor coat protein is free and (4) is appropriate if the carboxy terminus is free.

Similar constructions could be made with other filamentous phage. Pf3 is a well known filamentous phage that infects Pseudomonas aerugenosa cells that harbor an IncP-1 plasmid. The major coat protein of PF3 is unusual in having no signal peptide to direct its secretion. The sequence has charged residues ASP_7, ARG_{37}, LYS_{40}, and PHE_{44}-COO which is consistent with the amino terminus being exposed. Thus, to cause an IPBD to appear on the surface of Pf3, we construct a tripartite gene comprising:

1) a signal sequence known to cause secretion in P. aerugenosa (preferably known to cause secretion of IPBD) fused in-frame to,

2) a gene fragment encoding the IPBD sequence, fused in-frame to,

3) DNA encoding the mature Pf3 coat protein.

Optionally, DNA encoding a flexible linker of one to 10 amino acids and/or amino acids forming a recognition site for a specific protease (e.g., Factor Xa) is introduced between the ipbd gene fragment and the Pf3 coat-protein

gene. This tripartite gene is introduced into Pf3 so that it does not interfere with expression of any Pf3 genes. To reduce the possibility of genetic recombination, part (3) is designed to have numerous silent mutations relative to the wild-type gene. Once the signal sequence is cleaved off, the IPBD is in the periplasm and the mature coat protein acts as an anchor and phage-assembly signal. It does not matter that this fusion protein comes to rest in the lipid bilayer by a route different from the route followed by the wild-type coat protein.

As described in WO90/02809, other phage, such as bacteriophage ΦX174, large DNA phage such as λ or T4, and even RNA phage, may with suitable adaptations and modifications be used as GPs.

IV.C. Bacterial Cells as Genetic Packages:

One may choose any well-characterized bacterial strain which (1) may be grown in culture (2) may be engineered to display PBDs on its surface, and (3) is compatible with affinity selection.

Among bacterial cells, the preferred genetic packages are Salmonella typhimurium, Bacillus subtilis, Pseudomonas aeruginosa, Vibrio cholerae, Klebsiella pneumonia, Neisseria gonorrhoeae, Neisseria meningitidis, Bacteroides nodosus, Moraxella bovis, and especially Escherichia coli. The potential binding mini-protein may be expressed as an insert in a chimeric bacterial outer surface protein (OSP). All bacteria exhibit proteins on their outer surfaces. E. coli is the preferred bacterial GP and, for it, LamB is a preferred OSP.

While most bacterial proteins remain in the cytoplasm, others are transported to the periplasmic space (which lies between the plasma membrane and the cell wall of gram-negative bacteria), or are conveyed and anchored to the outer surface of the cell. Still others are exported (secreted) into the medium surrounding the cell. Those

characteristics of a protein that are recognized by a cell and that cause it to be transported out of the cytoplasm and displayed on the cell surface will be termed "outer-surface transport signals".

Gram-negative bacteria have outer-membrane proteins (OMP), that form a subset of OSPs. Many OMPs span the membrane one or more times. The signals that cause OMPs to localize in the outer membrane are encoded in the amino acid sequence of the mature protein. Outer membrane proteins of bacteria are initially expressed in a precursor form including a so-called signal peptide. The precursor protein is transported to the inner membrane, and the signal peptide moiety is extruded into the periplasmic space. There, it is cleaved off by a "signal peptidase", and the remaining "mature" protein can now enter the periplasm. Once there, other cellular mechanisms recognize structures in the mature protein which indicate that its proper place is on the outer membrane, and transport it to that location.

It is well known that the DNA coding for the leader or signal peptide from one protein may be attached to the DNA sequence coding for another protein, protein X, to form a chimeric gene whose expression causes protein X to appear free in the periplasm. The use of export-permissive bacterial strains (LISS85, STAD89) increases the probability that a signal-sequence-fusion will direct the desired protein to the cell surface.

OSP-IPBD fusion proteins need not fill a structural role in the outer membranes of Gram-negative bacteria because parts of the outer membranes are not highly ordered. For large OSPs there is likely to be one or more sites at which osp can be truncated and fused to ipbd such that cells expressing the fusion will display IPBDs on the cell surface. Fusions of fragments of omp genes with fragments of an x gene have led to X appearing on the outer membrane (CHAR88b,c, BENS84, CLEM81). When such fusions have been

made, we can design an osp-ipbd gene by substituting ipbd for x in the DNA sequence. Otherwise, a successful OMP-IPBD fusion is preferably sought by fusing fragments of the best omp to an ipbd, expressing the fused gene, and testing the resultant GPs for display-of-IPBD phenotype. We use the available data about the OMP to pick the point or points of fusion between omp and ipbd to maximize the likelihood that IPBD will be displayed. (Spacer DNA encoding flexible linkers, made, e.g., of GLY, SER, and ASN, may be placed between the osp- and ipbd-derived fragments to facilitate display.) Alternatively, we truncate osp at several sites or in a manner that produces osp fragments of variable length and fuse the osp fragments to ipbd; cells expressing the fusion are screened or selected which display IPBDs on the cell surface. Freudl et al. (FREU89) have shown that fragments of OSPs (such as OmpA) above a certain size are incorporated into the outer membrane. An additional alternative is to include short segments of random DNA in the fusion of omp fragments to ipbd and then screen or select the resulting variegated population for members exhibiting the display-of-IPBD phenotype.

In E. coli, the LamB protein is a well understood OSP and can be used. The E. coli LamB has been expressed in functional form in S. typhimurium, V. cholerae, and K. pneumonia, so that one could display a population of PBDs in any of these species as a fusion to E. coli LamB. K. pneumonia expresses a maltoporin similar to LamB (WEHM89) which could also be used. In P. aeruginosa, the D1 protein (a homologue of LamB) can be used (TRIA88).

LamB is transported to the outer membrane if a functional N-terminal sequence is present; further, the first 49 amino acids of the mature sequence are required for successful transport (BENS84). As with other OSPs, LamB of E. coli is synthesized with a typical signal-sequence which is subsequently removed. Homology between parts of LamB

protein and other outer membrane proteins OmpC, OmpF, and PhoE has been detected (NIKA84), including homology between LamB amino acids 39-49 and sequences of the other proteins. These subsequences may label the proteins for transport to the outer membrane.

The amino acid sequence of LamB is known (CLEM81), and a model has been developed of how it anchors itself to the outer membrane (Reviewed by, among others, BENZ88b). The location of its maltose and phage binding domains are also known (HEIN88). Using this information, one may identify several strategies by which a PBD insert may be incorporated into LamB to provide a chimeric OSP which displays the PBD on the bacterial outer membrane.

When the PBDs are to be displayed by a chimeric transmembrane protein like LamB, the PBD could be inserted into a loop normally found on the surface of the cell (cp. BECK83, MANO86). Alternatively, we may fuse a 5' segment of the *osp* gene to the *ipbd* gene fragment; the point of fusion is picked to correspond to a surface-exposed loop of the OSP and the carboxy terminal portions of the OSP are omitted. In LamB, it has been found that up to 60 amino acids may be inserted (CHAR88b,c) with display of the foreign epitope resulting; the structural features of OmpC, OmpA, OmpF, and PhoE are so similar that one expects similar behavior from these proteins.

It should be noted that while LamB may be characterized as a binding protein, it is used in the present invention to provide an OSTS; its binding domains are not variegated.

Other bacterial outer surface proteins, such as OmpA, OmpC, OmpF, PhoE, and pilin, may be used in place of LamB and its homologues. OmpA is of particular interest because it is very abundant and because homologues are known in a wide variety of gram-negative bacterial species. Baker *et al.* (BAKE87) review assembly of proteins into the outer membrane of *E. coli* and cite a topological model of OmpA

(VOGE86) that predicts that residues 19-32, 62-73, 105-118, and 147-158 are exposed on the cell surface. Insertion of a ipbd encoding fragment at about codon 111 or at about codon 152 is likely to cause the IPBD to be displayed on the cell surface. Concerning OmpA, see also MACI88 and MANO88. Porin Protein F of Pseudomonas aeruginosa has been cloned and has sequence homology to OmpA of E. coli (DUCH88). Although this homology is not sufficient to allow prediction of surface-exposed residues on Porin Protein F, the methods used to determine the topological model of OmpA may be applied to Porin Protein F. Works related to use of OmpA as an OSP include BECK80 and MACI88.

Misra and Benson (MISR88a, MISR88b) disclose a topological model of E. coli OmpC that predicts that, among others, residues GLY_{164} and LEU_{250} are exposed on the cell surface. Thus insertion of an ipbd gene fragment at about codon 164 or at about codon 250 of the E. coli ompC gene or at corresponding codons of the S. typhimurium ompC gene is likely to cause IPBD to appear on the cell surface. The ompC genes of other bacterial species may be used. Other works related to OmpC include CATR87 and CLIC88.

OmpF of E. coli is a very abundant OSP, $\geq 10^4$ copies/cell. Pages et al. (PAGE90) have published a model of OmpF indicating seven surface-exposed segments. Fusion of an ipbd gene fragment, either as an insert or to replace the 3' part of ompF, in one of the indicated regions is likely to produce a functional ompF::ipbd gene the expression of which leads to display of IPBD on the cell surface. In particular, fusion at about codon 111, 177, 217, or 245 should lead to a functional ompF::ipbd gene. Concerning OmpF, see also REID88b, PAGE88, BENS88, TOMM82, and SODE85.

Pilus proteins are of particular interest because piliated cells express many copies of these proteins and because several species (N. gonorrhoeae, P. aeruginosa,

Moraxella bovis, Bacteroides nodosus, and E. coli) express related pilins. Getzoff and coworkers (GETZ88, PARG87, SOME85) have constructed a model of the gonococcal pilus that predicts that the protein forms a four-helix bundle having structural similarities to tobacco mosaic virus protein and myohemerythrin. On this model, both the amino and carboxy termini of the protein are exposed. The amino terminus is methylated. Elleman (ELLE88) has reviewed pilins of Bacteroides nodosus and other species and serotype differences can be related to differences in the pilin protein and that most variation occurs in the C-terminal region. The amino-terminal portions of the pilin protein are highly conserved. Jennings et al. (JENN89) have grafted a fragment of foot-and-mouth disease virus (residues 144-159) into the B. nodosus type 4 fimbrial protein which is highly homologous to gonococcal pilin. They found that expression of the 3'-terminal fusion in P. aeruginosa led to a viable strain that makes detectable amounts of the fusion protein. Jennings et al. did not vary the foreign epitope nor did they suggest any variation. They inserted a GLY-GLY linker between the last pilin residue and the first residue of the foreign epitope to provide a "flexible linker". Thus a preferred place to attach an IPBD is the carboxy terminus. The exposed loops of the bundle could also be used, although the particular internal fusions tested by Jennings et al. (JENN89) appeared to be lethal in P. aeruginosa. Concerning pilin, see also MCKE85 and ORND85.

Judd (JUDD86, JUDD85) has investigated Protein IA of N. gonorrhoeae and found that the amino terminus is exposed; thus, one could attach an IPBD at or near the amino terminus of the mature P.IA as a means to display the IPBD on the N. gonorrhoeae surface.

A model of the topology of PhoE of E. coli has been disclosed by van der Ley et al. (VAND86). This model predicts eight loops that are exposed; insertion of an IPBD

into one of these loops is likely to lead to display of the IPBD on the surface of the cell. Residues 158, 201, 238, and 275 are preferred locations for insertion of and IPBD.

Other OSPs that could be used include E. coli BtuB, FepA, FhuA, IutA, FecA, and FhuE (GUDM89) which are receptors for nutrients usually found in low abundance. The genes of all these proteins have been sequenced, but topological models are not yet available. Gudmunsdottir et al. (GUDM89) have begun the construction of such a model for BtuB and FepA by showing that certain residues of BtuB face the periplasm and by determining the functionality of various BtuB::FepA fusions. Carmel et al. (CARM90) have reported work of a similar nature for FhuA. All Neisseria species express outer surface proteins for iron transport that have been identified and, in many cases, cloned. See also MORS87 and MORS88.

Many gram-negative bacteria express one or more phospholipases. E. coli phospholipase A, product of the pldA gene, has been cloned and sequenced by de Geus et al. (DEGE84). They found that the protein appears at the cell surface without any posttranslational processing. A ipbd gene fragment can be attached at either terminus or inserted at positions predicted to encode loops in the protein. That phospholipase A arrives on the outer surface without removal of a signal sequence does not prove that a PldA::IPBD fusion protein will also follow this route. Thus we might cause a PldA::IPBD or IPBD::PldA fusion to be secreted into the periplasm by addition of an appropriate signal sequence. Thus, in addition to simple binary fusion of an ipbd fragment to one terminus of pldA, the constructions:

1) ss::ipbd::pldA
2) ss::pldA::ipbd

should be tested. Once the PldA::IPBD protein is free in the periplasm it does not remember how it got there and the structural features of PldA that cause it to localize on the

outer surface will direct the fusion to the same destination.

IV.D. Bacterial Spores as Genetic Packages:

Bacterial spores have desirable properties as GP candidates. Spores are much more resistant than vegetative bacterial cells or phage to chemical and physical agents, and hence permit the use of a great variety of affinity selection conditions. Also, Bacillus spores neither actively metabolize nor alter the proteins on their surface. Bacillus spores, and more especially B. subtilis spores, are therefore the preferred sporoidal GPs. As discussed more fully in WO90/02809, a foreign binding domain may be introduced into an outer surface protein such as that encoded by the B. subtilis cotC or cotD genes.

It is generally preferable to use as the genetic package a cell, spore or virus for which an outer surface protein which can be engineered to display a IPBD has already been identified. However, as explained in WO90/02809, the present invention is not limited to such genetic packages, as an outer surface transport signal may be generated by variegation-and-selection techniques.

V.E Genetic Construction and Expression Considerations

The (i)pbd-osp gene may be: a) completely synthetic, b) a composite of natural and synthetic DNA, or c) a composite of natural DNA fragments. The important point is that the pbd segment be easily variegated so as to encode a multitudinous and diverse family of PBDs as previously described. A synthetic ipbd segment is preferred because it allows greatest control over placement of restriction sites. Primers complementary to regions abutting the osp-ipbd gene on its 3' flank and to parts of the osp-ipbd gene that are not to be varied are needed for sequencing.

The sequences of regulatory parts of the gene are taken from the sequences of natural regulatory elements: a) promoters, b) Shine-Dalgarno sequences, and c) trans-

criptional terminators. Regulatory elements could also be designed from knowledge of consensus sequences of natural regulatory regions. The sequences of these regulatory elements are connected to the coding regions; restriction sites are also inserted in or adjacent to the regulatory regions to allow convenient manipulation.

The essential function of the affinity separation is to separate GPs that bear PBDs (derived from IPBD) having high affinity for the target from GPs bearing PBDs having low affinity for the target. If the elution volume of a GP depends on the number of PBDs on the GP surface, then a GP bearing many PBDs with low affinity, $GP(PBD_w)$, might co-elute with a GP bearing fewer PBDs with high affinity, $GP(PBD_s)$. Regulation of the osp-pbd gene preferably is such that most packages display sufficient PBD to effect a good separation according to affinity. Use of a regulatable promoter to control the level of expression of the osp-pbd allows fine adjustment of the chromatographic behavior of the variegated population.

Induction of synthesis of engineered genes in vegetative bacterial cells has been exercised through the use of regulated promoters such as lacUV5, trpP, or tac (MANI82). The factors that regulate the quantity of protein synthesized are sufficiently well understood that a wide variety of heterologous proteins can now be produced in E. coli, B. subtilis and other host cells in at least moderate quantities (BETT88). Preferably, the promoter for the osp-ipbd gene is subject to regulation by a small chemical inducer. For example, the lac promoter and the hybrid trp-lac (tac) promoter are regulatable with isopropyl thiogalactoside (IPTG). The promoter for the constructed gene need not come from a natural osp gene; any regulatable bacterial promoter can be used. A non-leaky promoter is preferred.

The present invention is not limited to a single method of gene design. The osp-ipbd gene need not be synthesized in toto; parts of the gene may be obtained from nature. One may use any genetic engineering method to produce the correct gene fusion, so long as one can easily and accurately direct mutations to specific sites in the pbd DNA subsequence.

The coding portions of genes to be synthesized are designed at the protein level and then encoded in DNA. The ambiguity in the genetic code is exploited to allow optimal placement of restriction sites, to create various distributions of amino acids at variegated codons, to minimize the potential for recombination, and to reduce use of codons are poorly translated in the host cell.

V.F Structural Considerations

The design of the amino-acid sequence for the ipbd-osp gene to encode involves a number of structural considerations. The design is somewhat different for each type of GP. In bacteria, OSPs are not essential, so there is no requirement that the OSP domain of a fusion have any of its parental functions beyond lodging in the outer membrane.

It is desirable that the OSP not constrain the orientation of the PBD domain; this is not to be confused with lack of constraint within the PBD. Cwirla et al. (CWIR90), Scott and Smith (SCOT90), and Devlin et al. (DEVL90), have taught that variable residues in phage-displayed random peptides should be free of influence from the phage OSP. We teach that binding domains having a moderate to high degree of conformational constraint will exhibit higher specificity and that higher affinity is also possible. Thus, we prescribe picking codons for variegation that specify amino acids that will appear in a well-defined framework. The nature of the side groups is varied through a very wide range due to the combinatorial replacement of

multiple amino acids. The main chain conformations of most PBDs of a given class is very similar. The movement of the PBD relative to the OSP should not, however, be restricted. Thus it is often appropriate to include a flexible linker between the PBD and the OSP. Such flexible linkers can be taken from naturally occurring proteins known to have flexible regions. For example, the gIII protein of M13 contains glycine-rich regions thought to allow the amino-terminal domains a high degree of freedom. Such flexible linkers may also be designed. Segments of polypeptides that are rich in the amino acids GLY, ASN, SER, and ASP are likely to give rise to flexibility. Multiple glycines are particularly preferred.

When we choose to insert the PBD into a surface loop of an OSP such as LamB, OmpA, or M13 gIII protein, there are a few considerations that do not arise when PBD is joined to the end of an OSP. In these cases, the OSP exerts some constraining influence on the PBD; the ends of the PBD are held in more or less fixed positions. We could insert a highly varied DNA sequence into the osp gene at codons that encode a surface-exposed loop and select for cells that have a specific-binding phenotype. When the identified amino-acid sequence is synthesized (by any means), the constraint of the OSP is lost and the peptide is likely to have a much lower affinity for the target and a much lower specificity. Tan and Kaiser (TANN77) found that a synthetic model of BPTI containing all the amino acids of BPTI that contact trypsin has a K_d for trypsin $\sim 10^7$ higher than BPTI. Thus, it is strongly preferred that the varied amino acids be part of a PBD in which the structural constrains are supplied by the PBD.

It is known that the amino acids adjoining foreign epitopes inserted into LamB influence the immunological properties of these epitopes (VAND90). We expect that PBDs

inserted into loops of LamB, OmpA, or similar OSPs will be influenced by the amino acids of the loop and by the OSP in general. To obtain appropriate display of the PBD, it may be necessary to add one or more linker amino acids between the OSP and the PBD. Such linkers may be taken from natural proteins or designed on the basis of our knowledge of the structural behavior of amino acids. Sequences rich in GLY, SER, ASN, ASP, ARG, and THR are appropriate. One to five amino acids at either junction are likely to impart the desired degree of flexibility between the OSP and the PBD.

A preferred site for insertion of the ipbd gene into the phage osp gene is one in which: a) the IPBD folds into its original shape, b) the OSP domains fold into their original shapes, and c) there is no interference between the two domains.

If there is a model of the phage that indicates that either the amino or carboxy terminus of an OSP is exposed to solvent, then the exposed terminus of that mature OSP becomes the prime candidate for insertion of the ipbd gene. A low resolution 3D model suffices.

In the absence of a 3D structure, the amino and carboxy termini of the mature OSP are the best candidates for insertion of the ipbd gene. A functional fusion may require additional residues between the IPBD and OSP domains to avoid unwanted interactions between the domains. Random-sequence DNA or DNA coding for a specific sequence of a protein homologous to the IPBD or OSP, can be inserted between the osp fragment and the ipbd fragment if needed.

Fusion at a domain boundary within the OSP is also a good approach for obtaining a functional fusion. Smith exploited such a boundary when subcloning heterologous DNA into gene III of f1 (SMIT85).

The criteria for identifying OSP domains suitable for causing display of an IPBD are somewhat different from those used to identify and IPBD. When identifying an OSP, minimal

size is not so important because the OSP domain will not appear in the final binding molecule nor will we need to synthesize the gene repeatedly in each variegation round. The major design concerns are that: a) the OSP::IPBD fusion causes display of IPBD, b) the initial genetic construction be reasonably convenient, and c) the osp::ipbd gene be genetically stable and easily manipulated. There are several methods of identifying domains. Methods that rely on atomic coordinates have been reviewed by Janin and Chothia (JANI85). These methods use matrices of distances between α carbons (C_α), dividing planes (cf. ROSE85), or buried surface (RASH84). Chothia and collaborators have correlated the behavior of many natural proteins with domain structure (according to their definition). Rashin correctly predicted the stability of a domain comprising residues 206-316 of thermolysin (VITA84, RASH84).

Many researchers have used partial proteolysis and protein sequence analysis to isolate and identify stable domains. (See, for example, VITA84, POTE83, SCOT87a, and PABO79.) Pabo et al. used calorimetry as an indicator that the cI repressor from the coliphage λ contains two domains; they then used partial proteolysis to determine the location of the domain boundary.

If the only structural information available is the amino acid sequence of the candidate OSP, we can use the sequence to predict turns and loops. There is a high probability that some of the loops and turns will be correctly predicted (cf. Chou and Fasman, (CHOU74)); these locations are also candidates for insertion of the ipbd gene fragment.

In bacterial OSPs, the major considerations are: a) that the PBD is displayed, and b) that the chimeric protein not be toxic.

From topological models of OSPs, we can determine whether the amino or carboxy termini of the OSP is exposed. If so, then these are excellent choices for fusion of the osp fragment to the ipbd fragment.

The lamB gene has been sequenced and is available on a variety of plasmids (CLEM81, CHAR88a,b). Numerous fusions of fragments of lamB with a variety of other genes have been used to study export of proteins in E. coli. From various studies, Charbit et al. (CHAR88a,b) have proposed a model that specifies which residues of LamB are: a) embedded in the membrane, b) facing the periplasm, and c) facing the cell surface; we adopt the numbering of this model for amino acids in the mature protein. According to this model, several loops on the outer surface are defined, including: 1) residues 88 through 111, 2) residues 145 through 165, and 3) 236 through 251.

Consider a mini-protein embedded in LamB. For example, insertion of DNA encoding $G_1NXCX_5XXXCX_{10}SG_{12}$ between codons 153 and 154 of lamB is likely to lead to a wide variety of LamB derivatives being expressed on the surface of E. coli cells. G_1, N_2, S_{11}, and G_{12} are supplied to allow the mini-protein sufficient orientational freedom that is can interact optimally with the target. Using affinity enrichment (involving, for example, FACS via a fluorescently labeled target, perhaps through several rounds of enrichment), we might obtain a strain (named, for example, BEST) that expresses a particular LamB derivative that shows high affinity for the predetermined target. An octapeptide having the sequence of the inserted residues 3 through 10 from BEST is likely to have an affinity and specificity similar to that observed in BEST because the octapeptide has an internal structure that keeps the amino acids in a conformation that is quite similar in the LamB derivative and in the isolated mini-protein.

Fusing one or more new domains to a protein may make the ability of the new protein to be exported from the cell different from the ability of the parental protein. The signal peptide of the wild-type coat protein may function for authentic polypeptide but be unable to direct export of a fusion. To utilize the Sec-dependent pathway, one may need a different signal peptide. Thus, to express and display a chimeric BPTI/M13 gene VIII protein, we found it necessary to utilize a heterologous signal peptide (that of phoA).

GPs that display peptides having high affinity for the target may be quite difficult to elute from the target, particularly a multivalent target. (Bacteria that are bound very tightly can simply multiply in situ.) For phage, one can introduce a cleavage site for a specific protease, such as blood-clotting Factor Xa, into the fusion OSP protein so that the binding domain can be cleaved from the genetic package. Such cleavage has the advantage that all resulting phage have identical OSPs and therefore are equally infective, even if polypeptide-displaying phage can be eluted from the affinity matrix without cleavage. This step allows recovery of valuable genes which might otherwise be lost. To our knowledge, no one has disclosed or suggested using a specific protease as a means to recover an information-containing genetic package or of converting a population of phage that vary in infectivity into phage having identical infectivity.

IV.G. Synthesis of Gene Inserts

The present invention is not limited to any particular method or strategy of DNA synthesis or construction. Conventional DNA synthesizers may be used, with appropriate reagent modifications for production of variegated DNA (similar to that now used for production of mixed probes).

The osp-pbd genes may be created by inserting vgDNA into an existing parental gene, such as the osp-ipbd shown

to be displayable by a suitably transformed GP. The present invention is not limited to any particular method of introducing the vgDNA, e.g., cassette mutagenesis or single-stranded-oligonucleotide-directed mutagenesis

IV.H. Operative Cloning Vector

The operative cloning vector (OCV) is a replicable nucleic acid used to introduce the chimeric ipbd-osp or ipbd-osp gene into the genetic package. When the genetic package is a virus, it may serve as its own OCV. For cells and spores, the OCV may be a plasmid, a virus, a phagemid, or a chromosome.

IV.I. Transformation of cells:

When the GP is a cell, the population of GPs is created by transforming the cells with suitable OCVs. When the GP is a phage, the phage are genetically engineered and then transfected into host cells suitable for amplification. When the GP is a spore, cells capable of sporulation are transformed with the OCV while in a normal metabolic state, and then sporulation is induced so as to cause the OSP-PBDs to be displayed. The present invention is not limited to any one method of transforming cells with DNA.

The transformed cells are grown first under non-selective conditions that allow expression of plasmid genes and then selected to kill untransformed cells. Transformed cells are then induced to express the osp-pbd gene at the appropriate level of induction. The GPs carrying the IPBD or PBDs are then harvested by methods appropriate to the GP at hand, generally, centrifugation to pelletize GPs and resuspension of the pellets in sterile medium (cells) or buffer (spores or phage). They are then ready for verification that the display strategy was successful (where the GPs all display a "test" IPBD) or for affinity selection (where the GPs display a variety of different PBDs).

IV.J. Verification of Display Strategy:

The harvested packages are tested to determine whether the IPBD is present on the surface. In any tests of GPs for the presence of IPBD on the GP surface, any ions or cofactors known to be essential for the stability of IPBD or AfM(IPBD) are included at appropriate levels. The tests can be done, e.g., by a) by affinity labeling, b) enzymatically, c) spectrophotometrically, d) by affinity separation, or e) by affinity precipitation. The AfM(IPBD) in this step is one picked to have strong affinity (preferably, $K_d < 10^{-11}$ M) for the IPBD molecule and little or no affinity for the wtGP.

V. AFFINITY SELECTION OF TARGET-BINDING MUTANTS

V.A. Affinity Separation Technology, Generally

Affinity separation is used initially in the present invention to verify that the display system is working, i.e., that a chimeric outer surface protein has been expressed and transported to the surface of the genetic package and is oriented so that the inserted binding domain is accessible to target material. When used for this purpose, the binding domain is a known binding domain for a particular target and that target is the affinity molecule used in the affinity separation process. For example, a display system may be validated by using inserting DNA encoding BPTI into a gene encoding an outer surface protein of the genetic package of interest, and testing for binding to anhydrotrypsin, which is normally bound by BPTI.

If the genetic packages bind to the target, then we have confirmation that the corresponding binding domain is indeed displayed by the genetic package. Packages which display the binding domain (and thereby bind the target) are separated from those which do not.

Once the display system is validated, it is possible to use a variegated population of genetic packages which display a variety of different potential binding domains,

and use affinity separation technology to determine how well they bind to one or more targets. This target need not be one bound by a known binding domain which is parental to the displayed binding domains, i.e., one may select for binding to a new target.

The term "affinity separation means" includes, but is not limited to: a) affinity column chromatography, b) batch elution from an affinity matrix material, c) batch elution from an affinity material attached to a plate, d) fluorescence activated cell sorting, and e) electrophoresis in the presence of target material. "Affinity material" is used to mean a material with affinity for the material to be purified, called the "analyte". In most cases, the association of the affinity material and the analyte is reversible so that the analyte can be freed from the affinity material once the impurities are washed away.

V.B. Affinity Chromatography, Generally

Affinity column chromatography, batch elution from an affinity matrix material held in some container, and batch elution from a plate are very similar and hereinafter will be treated under "affinity chromatography."

If affinity chromatography is to be used, then:

1) the molecules of the target material must be of sufficient size and chemical reactivity to be applied to a solid support suitable for affinity separation,
2) after application to a matrix, the target material preferably does not react with water,
3) after application to a matrix, the target material preferably does not bind or degrade proteins in a non-specific way, and
4) the molecules of the target material must be sufficiently large that attaching the material to a matrix allows enough unaltered surface area (generally at least 500 Å^2, excluding the atom that is connected to the linker) for protein binding.

Affinity chromatography is the preferred separation means, but FACS, electrophoresis, or other means may also be used.

The present invention makes use of affinity separation of bacterial cells, or bacterial viruses (or other genetic packages) to enrich a population for those cells or viruses carrying genes that code for proteins with desirable binding properties.

V.C. Target Materials

The present invention may be used to select for binding domains which bind to one or more target materials, and/or fail to bind to one or more target materials. Specificity, of course, is the ability of a binding molecule to bind strongly to a limited set of target materials, while binding more weakly or not at all to another set of target materials from which the first set must be distinguished.

The target materials may be organic macromolecules, such as polypeptides, lipids, polynucleic acids, and polysaccharides, but are not so limited. The present invention is not, however, limited to any of the above-identified target materials. The only limitation is that the target material be suitable for affinity separation. Thus, almost any molecule that is stable in aqueous solvent may be used as a target.

Serine proteases such as human neutrophil elastase (HNE) are an especially interesting class of potential target materials. Serine proteases are ubiquitous in living organisms and play vital roles in processes such as: digestion, blood clotting, fibrinolysis, immune response, fertilization, and post-translational processing of peptide hormones. Although the role these enzymes play is vital, uncontrolled or inappropriate proteolytic activity can be very damaging.

V.D. Immobilization or Labeling of Target Material

For chromatography, FACS, or electrophoresis there may be a need to covalently link the target material to a second chemical entity. For chromatography the second entity is a matrix, for FACS the second entity is a fluorescent dye, and for electrophoresis the second entity is a strongly charged molecule. In many cases, no coupling is required because the target material already has the desired property of: a) immobility, b) fluorescence, or c) charge. In other cases, chemical or physical coupling is required.

It is not necessary that the actual target material be used in preparing the immobilized or labeled analogue that is to be used in affinity separation; rather, suitable reactive analogues of the target material may be more convenient. Target materials that do not have reactive functional groups may be immobilized by first creating a reactive functional group through the use of some powerful reagent, such as a halogen. In some cases, the reactive groups of the actual target material may occupy a part on the target molecule that is to be left undisturbed. In that case, additional functional groups may be introduced by synthetic chemistry.

Two very general methods of immobilization are widely used. The first is to biotinylate the compound of interest and then bind the biotinylated derivative to immobilized avidin. The second method is to generate antibodies to the target material, immobilize the antibodies by any of numerous methods, and then bind the target material to the immobilized antibodies. Use of antibodies is more appropriate for larger target materials; small targets (those comprising, for example, ten or fewer non-hydrogen atoms) may be so completely engulfed by an antibody that very little of the target is exposed in the target-antibody complex.

Non-covalent immobilization of hydrophobic molecules without resort to antibodies may also be used. A compound,

such as 2,3,3-trimethyldecane is blended with a matrix precursor, such as sodium alginate, and the mixture is extruded into a hardening solution. The resulting beads will have 2,3,3-trimethyldecane dispersed throughout and exposed on the surface.

Other immobilization methods depend on the presence of particular chemical functionalities. A polypeptide will present $-NH_2$ (N-terminal; Lysines), -COOH (C-terminal; Aspartic Acids; Glutamic Acids), -OH (Serines; Threonines; Tyrosines), and -SH (Cysteines). For the reactivity of amino acid side chains, see CREI84. A polysaccharide has free -OH groups, as does DNA, which has a sugar backbone.

Matrices suitable for use as support materials include polystyrene, glass, agarose and other chromatographic supports, and may be fabricated into beads, sheets, columns, wells, and other forms as desired.

Early in the selection process, relatively high concentrations of target materials may be applied to the matrix to facilitate binding; target concentrations may subsequently be reduced to select for higher affinity SBDs.

V.E. Elution of Lower Affinity PBD-Bearing Genetic Packages

The population of GPs is applied to an affinity matrix under conditions compatible with the intended use of the binding protein and the population is fractionated by passage of a gradient of some solute over the column. The process enriches for PBDs having affinity for the target and for which the affinity for the target is least affected by the eluants used. The enriched fractions are those containing viable GPs that elute from the column at greater concentration of the eluant.

The eluants preferably are capable of weakening noncovalent interactions between the displayed PBDs and the immobilized target material. Preferably, the eluants do not kill the genetic package; the genetic message corresponding

to successful mini-proteins is most conveniently amplified by reproducing the genetic package rather than by in vitro procedures such as PCR. The list of potential eluants includes salts (including Na+, NH_4+, Rb+, SO_4--, H_2PO_4-, citrate, K+, Li+, Cs+, HSO_4-, CO_3--, Ca++, Sr++, Cl-, PO_4---, HCO_3-, Mg++, Ba++, Br-, HPO_4-- and acetate), acid, heat, compounds known to bind the target, and soluble target material (or analogues thereof).

The uneluted genetic packages contain DNA encoding binding domains which have a sufficiently high affinity for the target material to resist the elution conditions. The DNA encoding such successful binding domains may be recovered in a variety of ways. Preferably, the bound genetic packages are simply eluted by means of a change in the elution conditions. Alternatively, one may culture the genetic package in situ, or extract the target-containing matrix with phenol (or other suitable solvent) and amplify the DNA by PCR or by recombinant DNA techniques. Additionally, if a site for a specific protease has been engineered into the display vector, the specific protease is used to cleave the binding domain from the GP.

Nonspecific binding to the matrix, etc., may be identified or reduced by techniques well known in the affinity separation art.

V.F. Recovery of packages:

Recovery of packages that display binding to an affinity column may be achieved in several ways, including:

1) collect fractions eluted from the column with a gradient as described above; fractions eluting later in the gradient contain GPs more enriched for genes encoding PBDs with high affinity for the column,
2) elute the column with the target material in soluble form,

3) flood the matrix with a nutritive medium and grow the desired packages in situ,

4) remove parts of the matrix and use them to inoculate growth medium,

5) chemically or enzymatically degrade the linkage holding the target to the matrix so that GPs still bound to target are eluted, or

6) degrade the packages and recover DNA with phenol or other suitable solvent; the recovered DNA is used to transform cells that regenerate GPs.

It is possible to utilize combinations of these methods. It should be remembered that what we want to recover from the affinity matrix is not the GPs per se, but the information in them. Recovery of viable GPs is very strongly preferred, but recovery of genetic material is essential. If cells, spores, or virions bind irreversibly to the matrix but are not killed, we can recover the information through in situ cell division, germination, or infection respectively. Proteolytic degradation of the packages and recovery of DNA is not preferred.

V.G. Amplifying the Enriched Packages

Viable GPs having the selected binding trait are amplified by culture in a suitable medium, or, in the case of phage, infection into a host so cultivated. If the GPs have been inactivated by the chromatography, the OCV carrying the osp-pbd gene are recovered from the GP, and introduced into a new, viable host.

V.H. Characterizing the Putative SBDs:

For one or more clonal isolates, we may subclone the sbd gene fragment, without the osp fragment, into an expression vector such that each SBD can be produced as a free protein. Physical measurements of the strength of binding may be made for each free SBD protein by any suitable method.

If we find that the binding is not yet sufficient, we decide which residues of the SBD (now a new PPBD) to vary next. If the binding is sufficient, then we now have a expression vector bearing a gene encoding the desired novel binding protein.

V.I. Joint selections:

One may modify the affinity separation of the method described to select a molecule that binds to material A but not to material B, or that binds to both A and B, either alternatively or simultaneously.

V.J. Engineering of Antagonists

It may be desirable to provide an antagonist to an enzyme or receptor. This may be achieved by making a molecule that prevents the natural substrate or agonist from reaching the active site. Molecules that bind directly to the active site may be either agonists or antagonists. Thus we adopt the following strategy. We consider enzymes and receptors together under the designation TER (Target Enzyme or Receptor).

For most TERs, there exist chemical inhibitors that block the active site. Usually, these chemicals are useful only as research tools due to highly toxicity. We make two affinity matrices: one with active TER and one with blocked TER. We make a variegated population of GP(PBD)s and select for SBPs that bind to both forms of the enzyme, thereby obtaining SDPs that do not bind to the active site. We expect that SBDs will be found that bind different places on the enzyme surface. Pairs of the sbd genes are fused with an intervening peptide segment. For example, if SBD-1 and SBD-2 are binding domains that show high affinity for the target enzyme and for which the binding is non-competitive, then the gene sbd-1::linker::sbd-2 encodes a two-domain protein that will show high affinity for the target. We make several fusions having a variety of SBDs and various

linkers. Such compounds have a reasonable probability of being an antagonist to the target enzyme.

VI. EXPLOITATION OF SUCCESSFUL BINDING DOMAINS AND CORRESPONDING DNAS

While the SBD may be produced by recombinant DNA techniques, an advantage inhering from the use of a mini-protein as an IPBD is that it is likely that the derived SBD will also behave like a mini-protein and will be obtainable by means of chemical synthesis. (The term "chemical synthesis", as used herein, includes the use of enzymatic agents in a cell-free environment.)

It is also to be understood that mini-proteins obtained by the method of the present invention may be taken as lead compounds for a series of homologues that contain non-naturally occurring amino acids and groups other than amino acids. For example, one could synthesize a series of homologues in which each member of the series has one amino acid replaced by its D enantiomer. One could also make homologues containing constituents such as β alanine, aminobutyric acid, 3-hydroxyproline, 2-Aminoadipic acid, N-ethylasperagine, norvaline, etc.; these would be tested for binding and other properties of interest, such as stability and toxicity.

Peptides may be chemically synthesized either in solution or on supports. Various combinations of stepwise synthesis and fragment condensation may be employed.

During synthesis, the amino acid side chains are protected to prevent branching. Several different protective groups are useful for the protection of the thiol groups of cysteines:

1) 4-methoxybenzyl (MBzl; Mob)(NISH82; ZAFA88), removable with HF;

2) acetamidomethyl (Acm) (NISH82; NISH86; BECK89c), removable with iodine; mercury ions (e.g., mercuric acetate); silver nitrate; and

3) S-para-methoxybenzyl (HOUG84).

Other thiol protective groups may be found in standard reference works such as Greene, PROTECTIVE GROUPS IN ORGANIC SYNTHESIS (1981).

Once the polypeptide chain has been synthesized, disulfide bonds must be formed. Possible oxidizing agents include air (HOUG84; NISH86), ferricyanide (NISH82; HOUG84), iodine (NISH82), and performic acid (HOUG84). Temperature, pH, solvent, and chaotropic chemicals may affect the course of the oxidation.

A large number of micro-proteins with a plurality of disulfide bonds have been chemically synthesized in biologically active form: conotoxin G1 (13AA, 4 Cys) (NISH-82); heat-stable enterotoxin ST (18AA, 6 Cys) (HOUG84); analogues of ST (BHAT86); Ω-conotoxin GVIA (27AA, 6Cys) (NISH86; RIVI87b); Ω-conotoxin MVIIA (27 AA, 6 Cys) (OLIV87b); α-conotoxin SI (13 AA, 4 Cys) (ZAFA88); μ-conotoxin IIIa (22AA, 6 Cys) (BECK89c, CRUZ89, HATA90). Sometimes, the polypeptide naturally folds so that the correct disulfide bonds are formed. Other times, it must be helped along by use of a differently removable protective group for each pair of cysteines.

The successful binding domains of the present invention may, alone or as part of a larger protein, be used for any purpose for which binding proteins are suited, including isolation or detection of target materials. In furtherance of this purpose, the novel binding proteins may be coupled directly or indirectly, covalently or noncovalently, to a label, carrier or support.

When used as a pharmaceutical, the novel binding proteins may be contained with suitable carriers or adjuvants.

EXAMPLE I

DESIGN AND MUTAGENESIS OF A CLASS 1 MICRO-PROTEIN

To obtain a library of binding domains that are conformationally constrained by a single disulfide, we insert DNA coding for the following family of micro-proteins into the gene coding for a suitable OSP.

$$X_1\text{-}C\text{-}X_2\text{-}X_3\text{-}X_4\text{-}X_5\text{-}C\text{-}X_6$$

Where ⌊——⌋ indicates disulfide bonding. Disulfides normally do not form between cysteines that are consecutive on the polypeptide chain. One or more of the residues indicated above as X_n will be varied extensively to obtain novel binding. There may be one or more amino acids that precede X_1 or follow X_6, however, the residues before X_1 or after X_6 will not be significantly constrained by the diagrammed disulfide bridge, and it is less advantageous to vary these remote, unbridged residues. The last X residue is connected to the OSP of the genetic package.

X_1, X_2, X_3, X_4, X_5, and X_6 can be varied independently; i.e. a different scheme of variegation could be used at each position. X_1 and X_6 are the least constrained residues and may be varied less than other positions.

X_1 and X_6 can be, for example, one of the amino acids [E, K, T, and A]; this set of amino acids is preferred because: a) the possibility of positively charged, negatively charged, and neutral amino acids is provided, b) these amino acids can be provided in 1:1:1:1 ratio via the codon RMG (R = equimolar A and G, M = equimolar A and C), and c) these amino acids allow proper processing by signal peptidases.

In a preferred embodiment, X_2, X_3, X_4 and X_5 are initially variegated by encoding each by the codon NNT,

which encodes the substitution set [F, S, Y, C, L, P, H, R, I, T, N, V, A, D, and G].

The advantages of the NNT over the NNK codon become increasingly apparent as the number of variegated codons increased. Tables 10 and 130 compare libraries in which six codons have been varied either by NNT or NNK codons. NNT encodes 15 different amino acids and only 16 DNA sequences. Thus, there are $1.139 \cdot 10^7$ amino-acid sequences, no stops, and only $1.678 \cdot 10^7$ DNA sequences. A library of 10^8 independent transformants will contain 99% of all possible sequences. The NNK library contains $6.4 \cdot 10^7$ sequences, but complete sampling requires a much larger number of independent transformants.

This sequence can be displayed as a fusion to the gene III protein of M13 using the native M13 gene III promoter and signal sequence. The sequence of M13 gene III protein, from residue 16 to 23, is $S_{16}HSAETVE_{23}$; signal peptidase-I cleaves after S_{18}. We replace this segment with

$$S_{16}GA_{18}AEGX_1CX_2X_3X_4X_5CX_6SYIEGRVIETVE.$$

Note that changing $H_{17}S_{18}$ to GA does not impare the phage for infectivity. It is useful to insert a bovine F.Xa recognition/cleavage site (YIEGR/VI) between the PBD and the mature III protein; this not only allows orientational freedom for the PBD, but also allows cleavage of the PBD from the GP.

A phage library in which X_1, X_2, X_5, and X_6 are encoded by NNT (allowing F, S, Y, C, L, P, H, R, V, T, N, V, A, D, & G) and in which X_3 and X_4 are encoded by NNG (allowing L, S, W, P, Q, R, M, T, K, V, A, E, and G) is named TN2. This library displays about 8.55×10^6 micro-proteins encoded by about 1.5×10^7 DNA sequences. NNG is used at the third and fourth variable positions (the central positions of the disulfide-closed loop) at least in part to avoid the possibility of cysteines at these positions.

Devlin, et al., screened 10^7 transformants, each of which could display one of 10^{12} random pentadecapeptides, for affinity with streptavidin, and found 20 streptavidin-binding phage isolates, with eight unique sequences ("A"-"I"). All contained HP; 15/20, HPQ; and 6/20, HPQF, though in different positions within the pentadecapeptide. The most frequently encountered isolates were D(5), I(4), and A(3), which entirely lacked cysteines. However, two positive isolates, "E"(1) and "F"(2), included a pair of cysteines positioned so that formation of a disulfide bond was possible. The sequences of these isolates is given in Table 820.

We recognized that our TN2 library should include a putative micro-protein, HPQ, similar enough to Devlin's "E" and "F" peptides to have the potential of exhibiting streptavidin-binding activity. HPQ comprises the AEG amino terminal sequence common to all members of the TN2 library, followed by the sequence PCHPQFCQ which has the potential for forming a disulfide bridge with a span of four, followed by a serine (S) and a bovine factor Xa recognition site (YIEGR/IV) (see Table 820). Pilot experiments showed that the binding of HPQ-bearing phage to streptavidin was comparable to that of Devlin's "F" isolate; both were marginally above background (1.7x). We therefore screened our TN2 library against immobilized streptavidin.

Streptavidin is available as free protein (Pierce) with a specific activity of 14.6 units per mg (1 unit will bind 1 μg of biotin). A stock solution of 1 mg per ml in PBS containing 0.01% azide is made. 100μL of StrAv stock is added to each 250 μL capacity well of Immulon (#4) plates and incubated overnight at 4°C. The stock is removed and replaced with 250 μL of PBS containing BSA at a concentration of 1 mg/mL and left at 4°C for a further 1 hour. Prior to use in a phage binding assay the wells are

washed rapidly 5 times with 250 μL of PBS containing 0.1% Tween.

To each StrAv-coated well is added 100 μL of binding buffer (PBS with 1 mg per mL BSA) containing a known quantity of phage (10^{11} pfu's of the TN2 library). Incubation proceeds for 1 hr at room temperature followed by removal of the non-bound phage and 10 rapid washes with PBS 0.1% Tween, then further washed with citrate buffers of pH 7, 6 and 5 to remove non-specific binding. The bound phage are eluted with 250 μL of pH2 citrate buffer containing 1 mg per mL BSA and neutralization with 60 μL of 1M tris pH 8. The eluate was used to infect bacterial cells which generated a new phage stock to be used for a further round of binding, washing and elution. The enhancement cycles were repeated two more times (three in total) after which time a number of individual phage were sequenced and tested as clonal isolates. The number of phage present in each step is determined as plaque forming units (pfu's) following appropriate dilutions and plating in a lawn of F' containing *E. coli*.

Table 838 shows the peptide sequences found to bind to StrAv and their frequency in the random picks taken from the final (round 3) phage pool.

The intercysteine segment of all of the putative micro-proteins examined contained the HPQF motif. The variable residue before the first cysteine could have contained any of {F,S,Y,C,L,P,H,R,I,T,N,V,A,D,G}; the residues selected were {Y,H,L,D,N} while phage HPQ has P. The variable residue after the second cysteine also could have had {F,S,Y,C,L,P,H,R,I,T,N,V,A,D,G}; the residues selected were {P,S,G,R,V} while phage HPQ has Q. The relatively poor binding of phage HPQ could be due to P_4 or to Q_{13} or both.

In a control experiment, the TN2 library was screened in an identical manner to that shown above but with the target protein being the blocking agent BSA. Following three rounds of binding, elution, and amplification, sixteen random phage plaques were picked and sequenced. Half of the clones demonstrated a lack of insert (8/16), the other half had the sequences shown in Table 839. There is no consensus for this collection.

We have displayed a related micro-protein, HPQ6, on phage. It is identical to HPQ except for the replacement of CHPQFC with CHPQFPRC (see Table 820). When displayed, HPQ6 had a substantially stronger affinity for streptavidin than either HPQ or Devlin's "F" isolate. (Devlin's "E" isolate was not studied.) Treatment with dithiothreitol (DTT) markedly reduced the binding of HPQ6 phage (but not control phage) to streptavidin, suggesting that the presence of a disulfide bridge within the displayed peptide was required for good binding. In view of the results of the screening of the TN2 library, it is likely that the binding of phage HPQ6 could be further improved by changing P_4 to one of {Y,H,L,D,N} and/or changing Q_{13} to one of {P,S,G,R,V}.

EXAMPLE II

A CYS::HELIX::TURN::STRAND::CYS UNIT

The parental Class 2 micro-protein may be a naturally-occurring Class 2 micro-protein. It may also be a domain of a larger protein whose structure satisfies or may be modified so as to satisfy the criteria of a class 2 micro-protein. The modification may be a simple one, such as the introduction of a cysteine (or a pair of cysteines) into the base of a hairpin structure so that the hairpin may be closed off with a disulfide bond, or a more elaborate one, so as the modification of intermediate residues so as to achieve the hairpin structure. The parental class 2 micro-

protein may also be a composite of structures from two or more naturally-occurring proteins, e.g., an α helix of one protein and a β strand of a second protein.

One micro-protein motif of potential use comprises a disulfide loop enclosing a helix, a turn, and a return strand. Such a structure could be designed or it could be obtained from a protein of known 3D structure. Scorpion neurotoxin, variant 3, (ALMA83a, ALMA83b) (hereafter ScorpTx) contains a structure diagrammed in Figure 1 that comprises a helix (residues N22 through N33), a turn (residues 33 through 35), and a return strand (residues 36 through 41). ScorpTx contains disulfides that join residues 12-65, 16-41, 25-46, and 29-48. CYS_{25} and CYS_{41} are quite close and could be joined by a disulfide without deranging the main chain. Figure 1 shows CYS_{25} joined to CYS_{41}. In addition, CYS_{29} has been changed to GLN. It is expected that a disulfide will form between 25 and 41 and that the helix shown will form; we know that the amino-acid sequence shown is highly compatible with this structure. The presence of GLY_{35}, GLY_{36}, and GLY_{39} give the turn and extended strand sufficient flexibility to accommodate any changes needed around CYS_{41} to form the disulfide.

From examination of this structure (as found in entry 1SN3 of the Brookhaven Protein Data Bank), we see that the following sets of residues would be preferred for variegation:

SET 1

Residue	Codon	Allowed amino acids	Naa/Ndna
1) T_{27}	NNG	$L^2R^2MVSPTAQKEWG.$	13/15
2) E_{28}	VHG	LMVPTAGKE	9/9
3) A_{31}	VHG	LMVPTAGKE	9/9
4) K_{32}	VHG	LMVPTAGKE	9/9
5) G_{24}	NNG	$L^2R^2MVSPTAQKEWG.$	13/15
6) E_{23}	VHG	LMVPTAGKE	9/9
7) Q_{34}	VAS	HQNKED	6/6

Note: Exponents on amino acids indicate multiplicity of codons.

Positions 27, 28, 31, 32, 24, and 23 comprise one face of the helix. At each of these locations we have picked a variegating codon that a) includes the parental amino acid, b) includes a set of residues having a predominance of helix favoring residues, c) provides for a wide variety of amino acids, and d) leads to as even a distribution as possible. Position 34 is part of a turn. The side group of residue 34 could interact with molecules that contact the side groups of resideus 27, 28, 31, 32, 24, and 23. Thus we allow variegation here and provide amino acids that are compatible with turns. The variegation shown leads to $6.65 \cdot 10^6$ amino acid sequences encoded by $8.85 \cdot 10^6$ DNA sequences.

SET 2

Residue	Codon	Allowed amino acids	Naa/Ndna
1) D_{26}	VHS	$L^2IMV^2P^2T^2A^2HQNKDE$	13/18
2) T_{27}	NNG	$L^2R^2MVSPTAQKEWG.$	13/15
3) K_{30}	VHG	KEQPTALMV	9/9
4) A_{31}	VHG	KEQPTALMV	9/9
5) K_{32}	VHG	LMVPTAGKE	9/9
6) S_{37}	RRT	SNDG	4/4
7) Y_{38}	NHT	YSFHPLNTIDAV	9/9

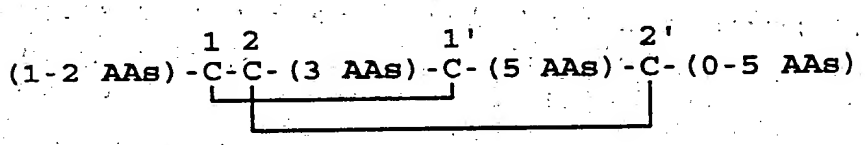
Positions 26, 27, 30, 31, and 32 are variegated so a to enhance helix-favoring amino acids in the population. Residues 37 and 38 are in the return strand so that we pick different variegation codons. This variegation allows $4.43 \cdot 10^6$ amino-acid sequences and $7.08 \cdot 10^6$ DNA sequences. Thus a library that embodies this scheme can be sampled very efficiently.

EXAMPLE III

DESIGN AND MUTAGENESIS OF CLASS 3 MICRO-PROTEIN

Two Disulfide Bond Parental Micro-Proteins

Micro-proteins with two disulfide bonds may be modelled after the α-conotoxins, e.g., GI, GIA, GII, MI, and SI. These have the following conserved structure:



Hashimoto et al. (HASH85) reported synthesis of twenty-four analogues of α conotoxins GI, GII, and MI. Using the numbering scheme for GI (CYS at positions 2, 3, 7, and 13), Hashimoto et al. reported alterations at 4, 8, 10, and 12 that allows the proteins to be toxic. Almquist et al. (ALMQ89) synthesized [des-GLU_1] α Conotoxin GI and twenty analogues. They found that substituting GLY for PRO_5 gave rise to two isomers, perhaps related to different disulfide bonding. They found a number of substitutions at residues 8 through 11 that allowed the protein to be toxic. Zafaralla et al. (ZAFA88) found that substituting PRO at position 9 gives an active protein. Each of the groups cited used only in vivo toxicity as an assay for the activity. From such studies, one can infer that an active protein has the

parental 3D structure, but one can not infer that an inactive protein lacks the parental 3D structure.

Pardi et al. (PARD89) determined the 3D structure of α Conotoxin GI obtained from venom by NMR. Kobayashi et al. (KOBA89) have reported a 3D structure of synthetic α Conotoxin GI from NMR data which agrees with that of PARD89. We refer to Figure 5 of Pardi et al..

Residue GLU_1 is known to accomodate GLU, ARG, and ILE in known analogues or homologues. A preferred variegation codon is NNG that allows the set of amino acids [L^2R^2MVSPTA-QKEWG<stop>]. From Figure 5 of Pardi et al. we see that the side group of GLU_1 projects into the same region as the strand comprising residues 9 through 12. Residues 2 and 3 are cysteines and are not to be varied. The side group of residue 4 points away from residues 9 through 12; thus we defer varying this residue until a later round. PRO_5 may be needed to cause the correct disulfides to form; when GLY was substituted here the peptide folded into two forms, neither of which is toxic. It is allowed to vary PRO_5, but not perferred in the first round.

No substitutions at ALA_6 have been reported. A preferred variegation codon is RMG which gives rise to ALA, THR, LYS, and GLU (small hydrophobic, small hydrophilic, positive, and negative). CYS_7 is not varied. We prefer to leave GLY_8 as is, although a homologous protein having ALA_8 is toxic. Homologous proteins having various amino acids at position 9 are toxic; thus, we use an NNT variegation codon which allows FS^2YCLPHRITNVADG. We use NNT at positions 10, 11, and 12 as well. At position 14, following the fourth CYS, we allow ALA, THR, LYS, or GLU (via an RMG codon). This variegation allows $1.053 \cdot 10^7$ anino-acid sequences, encoded by $1.68 \cdot 10^7$ DNA sequences. Libraries having $2.0 \cdot 10^7$, $3.0 \cdot 10^7$, and $5.0 \cdot 10^7$ independent transformants will, respectively, display ~70%, ~83%, and ~95% of the allowed

sequences. Other variegations are also appropriate. Concerning α conotoxins, see, inter alia, ALMQ89, CRUZ85, GRAY83, GRAY84, and PARD89.

The parental micro-protein may instead be one of the proteins designated "Hybrid-I" and "Hybrid-II" by Pease et al. (PEAS90); cf. Figure 4 of PEAS90. One preferred set of residues to vary for either protein consists of:

Parental Amino acid	Variegated Codon	Allowed Amino acids	AA seqs/ DNA seqs
A5	RVT	ADGTNS	6/6
P6	VYT	PTALIV	6/6
E7	RRS	EDNKSRG[2]	7/8
T8	VHG	TPALMVQKE	9/9
A9	VHG	ATPLMVQKE	9/9
A10	RMG	AEKT	4/4
K12	VHG	KQETPALMV	9/9
Q16	NNG	L[2]R[2]S.WPQMTKVAEG	13/15

This provides $9.55 \cdot 10^6$ amino-acid sequences encoded by $1.26 \cdot 10^7$ DNA sequences. A library comprising $5.0 \cdot 10^7$ transformants allows expression of 98.2% of all possible sequences. At each position, the parental amino acid is allowed.

At position 5 we provide amino acids that are compatible with a turn. At position 6 we allow ILE and VAL because they have branched β carbons and make the chain ridged. At position 7 we allow ASP, ASN, and SER that often appear at the amino termini of helices. At positions 8 and 9 we allow several helix-favoring amino acids (ALA, LEU, MET, GLN, GLU, and LYS) that have differing charges and hydrophobicities because these are part of the helix proper. Position 10 is further around the edge of the helix, so we allow a smaller set (ALA, THR, LYS, and GLU). This set not only includes 3 helix-favoring amino acids plus THR that is well tolerated but also allows positive, negative, and neutral hydrophilic.

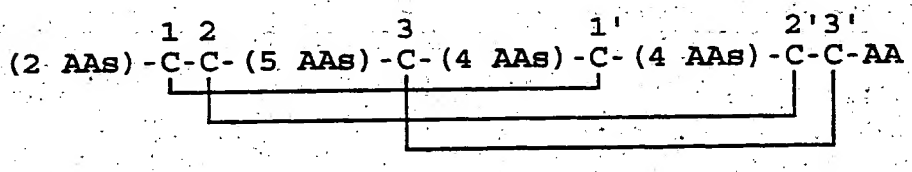
The side groups of 12 and 16 project into the same region as the residues already recited. At these positions we allow a wide variety of amino acids with a bias toward helix-favoring amino acids.

The parental micro-protein may instead be a polypeptide composed of residues 9-24 and 31-40 of aprotinin and possessing two disulfides (Cys9-Cys22 and Cys14-Cys38). Such a polypeptide would have the same disulfide bond topology as α-conotoxin, and its two bridges would have spans of 12 and 17, respectively.

Residues 23, 24 and 31 are variegated to encode the amino acid residue set [G,S,R,D,N,H,P,T,A] so that a sequence that favors a turn of the necessary geometry is found. We use trypsin or anhydrotrypsin as the affinity molucule to enrich for GPs that display a micro-protein that folds into a stable structure similar to BPTI in the P1 region.

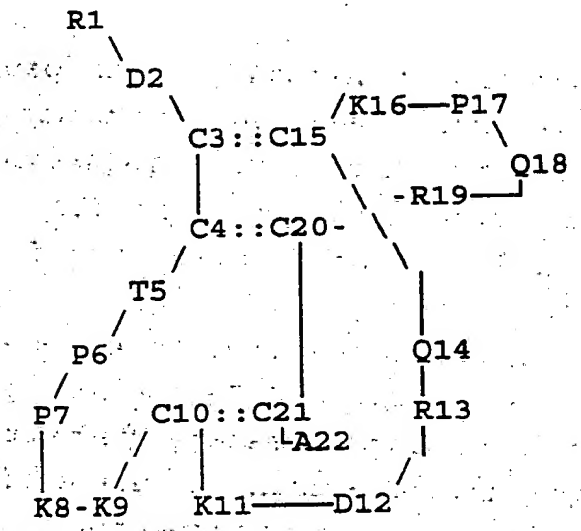
Three Disulfide Bond Parental Micro-Proteins

The cone snails (Conus) produce venoms (conotoxins) which are 10-30 amino acids in length and exceptionally rich in disulfide bonds. They are therefore archetypal micro-proteins. Novel micro-proteins with three disulfide bonds may be modelled after the μ-(GIIIA, GIIIB, GIIIC) or Ω-(GVIA, GVIB, GVIC, GVIIA, GVIIB, MVIIA, MVIIB, etc.) conotoxins. The μ-conotoxins have the following conserved structure:



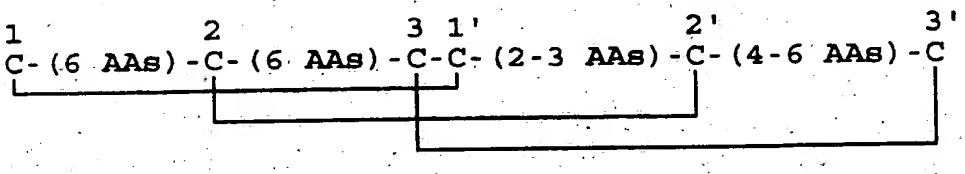
No 3D structure of a μ-conotoxin has been published. Hidaka et al. (HIDA90) have established the connectivity of

the disulfides. The following diagram depicts geographutoxin I (also known as μ-conotoxin GIIIA).



The connection from R19 to C20 could go over or under the strand from Q14 to C15. One preferred form of variegation is to vary the residues in one loop. Because the longest loop contains only five amino acids, it is appropriate to also vary the residues connected to the cysteines that form the loop. For example, we might vary residues 5 through 9 plus 2, 11, 19, and 22. Another useful variegation would be to vary residues 11-14 and 16-19, each through eight amino acids. Concerning μ conotoxins, see BECK89b, BECK89c, CRUZ89, and HIDA90.

The Ω-conotoxins may be represented as follows:



The King Kong peptide has the same disulfide arrangement as the Ω-conotoxins but a different biological activity. Woodward et al. (WOOD90) report the sequences of three

homologuous proteins from C. textile. Within the mature toxin domain, only the cysteines are conserved. The spacing of the cysteines is exactly conserved, but no other position has the same amino acid in all three sequences and only a few positions show even pair-wise matches. Thus we conclude that all positions (except the cysteines) may be substituted freely with a high probability that a stable disulfide structure will form. Concerning Ω conotoxins, see HILL89 and SUNX87.

Another micro-protein which may be used as a parental binding domain is the Cucurbita maxima trypsin inhibitor I (CMTI-I); CMTI-III is also appropriate. They are members of the squash family of serine protease inhibitors, which also includes inhibitors from summer squash, zucchini, and cucumbers (WIEC85). McWherter et al. (MCWH89) describe synthetic sequence-variants of the squash-seed protease inhibitors that have affinity for human leukocyte elastase and cathepsin G. Of course, any member of this family might be used.

CMTI-I is one of the smallest proteins known, comprising only 29 amino acids held in a fixed comformation by three disulfide bonds. The structure has been studied by Bode and colleagues using both X-ray diffraction (BODE89) and NMR (HOLA89a,b). CMTI-I is of ellipsoidal shape; it lacks helices or β-sheets, but consists of turns and connecting short polypeptide stretches. The disulfide pairing is Cys3-Cys20, Cys10-Cys22 and Cys16-Cys28. In the CMTI-I:trypsin complex studied by Bode et al., 13 of the 29 inhibitor residues are in direct contact with trypsin; most of them are in the primary binding segment Val2(P4)-Glu9 (P4') which contains the reactive site bond Arg5(P1)-Ile6 and is in a conformation observed also for other serine proteinase inhibitors.

CMTI-I has a K_i for trypsin of ~$1.5 \cdot 10^{-12}$ M. McWherter et al. suggested substitution of "moderately bulky hydrophobic groups" at P1 to confer HLE specificity. They found that a wider set of residues (VAL, ILE, LEU, ALA, PHE, MET, and GLY) gave detectable binding to HLE. For cathepsin G, they expected bulky (especially aromatic) side groups to be strongly preferred. They found that PHE, LEU, MET, and ALA were functional by their criteria; they did not test TRP, TYR, or HIS. (Note that ALA has the second smallest side group available.)

A preferred initial variegation strategy would be to vary some or all of the residues ARG_1, VAL_2, PRO_4, ARG_5, ILE_6, LEU_7, MET_8, GLU_9, LYS_{11}, HIS_{25}, GLY_{26}, TYR_{27}, and GLY_{29}. If the target were HNE, for example, one could synthesize DNA embodying the following possibilities:

Parental	vg Codon	Allowed amino acids	#AA seqs/ #DNA seqs
ARG_1	VNT	RSLPHITNVADG	12/12
VAL_2	NWT	VILFYHND	8/8
PRO_4	VYT	PLTIAV	6/6
ARG_5	VNT	RSLPHITNVADG	12/12
ILE6	NNK	all 20	20/31
LEU_7	VWG	LQMKVE	6/6
TYR_{27}	NAS	YHQNKDE.	7/8

This allows about $5.81 \cdot 10^6$ amino-acid sequences encoded by about $1.03 \cdot 10^7$ DNA sequences. A library comprising $5.0 \cdot 10^7$ independent transformants would give ~99% of the possible sequences. Other variegation schemes could also be used.

Other inhibitors of this family include:
Trypsin inhibitor I from Citrullus vulgaris (OTLE87),
Trypsin inhibitor II from Bryonia dioica (OTLE87),
Trypsin inhibitor I from Cucurbita maxima (in OTLE87),
trypsin inhibitor III from Cucurbita maxima (in OTLE87),
trypsin inhibitor IV from Cucurbita maxima (in OTLE87),

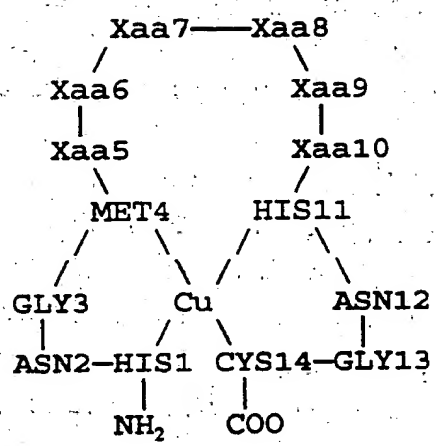
trypsin inhibitor II from Cucurbita pepo (in OTLE87), trypsin inhibitor III from Cucurbita pepo (in OTLE87), trypsin inhibitor IIb from Cucumis sativus (in OTLE87), trypsin inhibitor IV from Cucumis sativus (in OTLE87), trypsin inhibitor II from Ecballium elaterium (FAVE89), and inhibitor CM-1 from Momordica repens (in OTLE87).

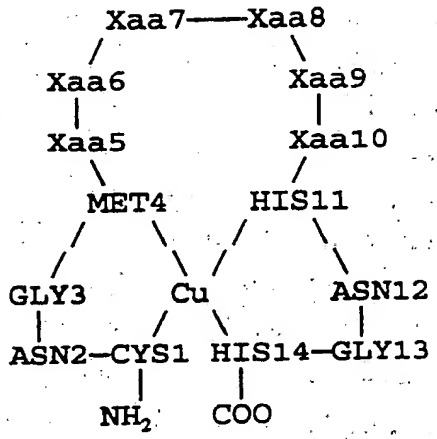
Another micro-protein that may be used as an initial potential binding domain is the heat-stable enterotoxins derived from some enterotoxogenic E. coli, Citrobacter freundii, and other bacteria (GUAR89). These micro-proteins are known to be secreted from E. coli and are extremely stable. Works related to synthesis, cloning, expression and properties of these proteins include: BHAT86, SEKI85, SHIM87, TAKA85, TAKE90, THOM85a,b, YOSH85, DALL90, DWAR89, GARI87, GUZM89, GUZM90, HOUG84, KUBO89, KUPE90, OKAM87, OKAM88, and OKAM90.

EXAMPLE IV

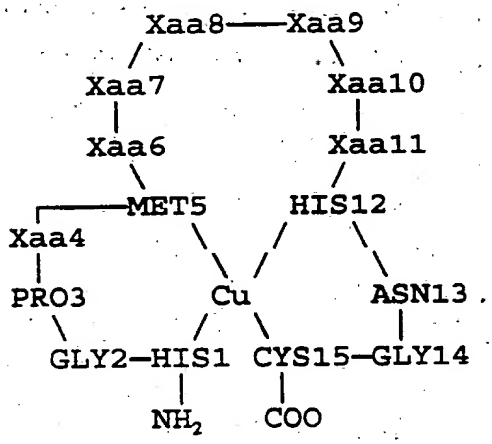
A MINI-PROTEIN HAVING A CROSS-LINK CONSISTING OF CU(II), ONE CYSTEINE, TWO HISTIDINES, AND ONE METHIONINE.

Sequences such as
HIS-ASN-GLY-MET-Xaa-Xaa-Xaa-Xaa-Xaa-Xaa-HIS-ASN-GLY-CYS and
CYS-ASN-GLY-MET-Xaa-Xaa-Xaa-Xaa-Xaa-Xaa-HIS-ASN-GLY-HISare likely to combine with Cu(II) to form structures as shown in the diagram:





Other arrangements of HIS, MET, HIS, and CYS along the chain are also likely to form similar structures. The amino acids ASN-GLY at positions 2 and 3 and at positions 12 and 13 give the amino acids that carry the metal-binding ligands enough flexibility for them to come together and bind the metal. Other connecting sequences may be used, e.g. GLY-ASN, SER-GLY, GLY-PRO, GLY-PRO-GLY, or PRO-GLY-ASN could be used. It is also possible to vary one or more residues in the loops that join the first and second or the third and fourth metal-binding residues. For example,



is likely to form the diagrammed structure for a wide variety of amino acids at Xaa4. It is expected that the side groups of Xaa4 and Xaa6 will be close together and on the surface of the mini-protein.

The variable amino acids are held so that they have limited flexibility. This cross-linkage has some differences from the disulfide linkage. The separation between $C_{\alpha 4}$ and $C_{\alpha 11}$ is greater than the separation of the C_αs of a cystine. In addition, the interaction of residues 1 through 4 and 11 through 14 with the metal ion are expected to limit the motion of residues 5 through 10 more than a disulfide between rsidues 4 and 11. A single disulfide bond exerts strong distance constrains on the α carbons of the joined residues, but very little directional constraint on, for example, the vector from N to C in the main-chain.

For the desired sequence, the side groups of residues 5 through 10 can form specific interactions with the target. Other numbers of variable amino acids, for example, 4, 5, 7, or 3, are appropriate. Larger spans may be used when the enclosed sequence contains segments having a high potential to form α helices or other secondary structure that limits the conformational freedom of the polypeptide main chain. Whereas a mini-protein having four CYSs could form three distinct pairings, a mini-protein having two HISs, one MET, and one CYS can form only two distinct complexes with Cu. These two structures are related by mirror symmetry through the Cu. Because the two HISs are distinguishable, the structures are different.

When such metal-containing mini-proteins are displayed on filamentous phage, the cells that produce the phage can be grown in the presence of the appropriate metal ion, or the phage can be exposed to the metal only after they are separated from the cells.

EXAMPLE V

A MINI-PROTEIN HAVING A CROSS-LINK CONSISTING OF ZN(II) AND FOUR CYSTEINES

A cross link similar to the one shown in Example XV is exemplified by the Zinc-finger proteins (GIBS88, GAUS87, PARR88, FRAN87, CHOW87, HARD90). One family of Zinc-fingers has two CYS and two HIS residues in conserved positions that bind Zn_{++} (PARR88, FRAN87, CHOW87, EVAN88, BERG88, CHAV88). Gibson et al. (GIBS88) review a number of sequences thought to form zinc-fingers and propose a three-dimensional model for these compounds. Most of these sequences have two CYS and two HIS residues in conserved positions, but some have three CYS and one HIS residue. Gauss et al. (GAUS87) also report a zinc-finger protein having three CYS and one HIS residues that bind zinc. Hard et al. (HARD90) report the 3D structure of a protein that comprises two zinc-fingers, each of which has four CYS residues. All of these zinc-binding proteins are stable in the reducing intracellular environment.

One preferred example of a CYS::zinc cross linked mini-protein comprises residues 440 to 461 of the sequence shown in Figure 1 of HARD90. The resiudes 444 through 456 may be variegated. One such variegation is as follows:

Parental	Allowed	#AA / #DNA
SER444	SER, ALA	2 / 2
ASP445	ASP, ASN, GLU, LYS	4 / 4
GLU446	GLU, LYS, GLN	3 / 3
ALA447	ALA, THR, GLY, SER	4 / 4
SER448	SER, ALA	2 / 2
GLY449	GLY, SER, ASN, ASP	4 / 4
CYS450	CYS, PHE, ARG, LEU	4 / 4
HIS451	HIS, GLN, ASN, LYS, ASP, GLU	6 / 6
TYR452	TYR, PHE, HIS, LEU	4 / 4
GLY453	GLY, SER, ASN, ASP	4 / 4
VAL454	VAL, ALA, ASP, GLY, SER, ASN, THR, ILE	8 / 8
LEU455	LEU, HIS, ASP, VAL	4 / 4
THR456	THR, ILE, ASN, SER	4 / 4

This leads to $3.77 \cdot 10^7$ DNA sequences that encode the same number of amino-acid sequences. A library having $1.0 \cdot 10^8$ independent transformants will display 93% of the allowed sequences; $2.0 \cdot 10^8$ independent transformants will display 99.5% of allowed sequences.

Table 2: Preferred Outer-Surface Proteins

Genetic Package preference	Preferred Outer-Surface Protein	Reason for
M13	coat protein (gpVIII)	a) exposed amino terminus, b) predictable post-translational processing, c) numerous copies in virion. d) fusion data available
	gp III	a) fusion data available. b) amino terminus exposed. c) working example available.
PhiX174	G protein	a) known to be on virion exterior, b) small enough that the G-ipbd gene can replace H gene.
E. coli	LamB	a) fusion data available, b) non-essential.
	OmpC	a) topological model b) non-essential; abundant
	OmpA	a) topological model b) non-essential; abundant c) homologues in other genera

100

O m p F

a) topological model
b) non-essential; abundant

PhoEa) topological model
b) non-essential; abundant
c) inducible

B. subtilis	CotC	a) no post-translational spores processing, b) distinctive sdequence that causes protein to localize in spore coat, c) non-essential.
	CotD	Same as for CotC.

Table 10: Abundances obtained from various vgCodons

A. Optimized fxS Codon, Restrained by [D]+[E] = [K]+[R]

	T	C	A	G	
1	.26	.18	.26	.30	f
2	.22	.16	.40	.22	x
3	.5	.0	.0	.5	S

Amino acid	Abundance	Amino acid	Abundance
A	4.80%	C	2.86%
D	6.00%	E	6.00%
F	2.86%	G	6.60%
H	3.60%	I	2.86%
K	5.20%	L	6.82%
M	2.86%	N	5.20%
P	2.88%	Q	3.60%
R	6.82%	S	7.02% mfaa
T	4.16%	V	6.60%
W	2.86% lfaa	Y	5.20%
stop	5.20%		

[D] + [E] = [K] + [R] = .12

ratio = Abun(W)/Abun(S) = 0.4074

i	$(1/ratio)^i$	$(ratio)^i$	stop-free
1	2.454	.4074	.9480
2	6.025	.1660	.8987
3	14.788	.0676	.8520
4	36.298	.0275	.8077
5	89.095	.0112	.7657
6	218.7	$4.57 \cdot 10^{-3}$	.7258
7	536.8	$1.86 \cdot 10^{-3}$	.6881

Table 10: Abundances obtained from various vgCodon (continued)

B. Unrestrained, optimized

	T	C	A	G
1	.27	.19	.27	.27
2	.21	.15	.43	.21
3	.5	.0	.0	.5

Amino acid	Abundance	Amino acid	Abundance
A	4.05%	C	2.84%
D	5.81%	E	5.81%
F	2.84%	G	5.67%
H	4.08%	I	2.84%
K	5.81%	L	6.83%
M	2.84%	N	5.81%
P	2.85%	Q	4.08%
R	6.83%	S	6.89% mfaa
T	4.05%	V	5.67%
W	2.84% lfaa	Y	5.81%
stop	5.81%		

[D] + [E] = 0.1162 [K] + [R] = 0.1264

ratio = Abun(W)/Abun(S) = 0.41176

j	$(1/ratio)^j$	$(ratio)^j$	stop-free
1	2.4286	.41176	.9419
2	5.8981	.16955	.8872
3	14.3241	.06981	.8356
4	34.7875	.02875	.7871
5	84.4849	.011836	.74135
6	205.180	.004874	.69828
7	498.3	$2.007 \cdot 10^{-3}$	.6577

Table 10: Abundances obtained from various vgCodon (continued)

C. Optimized NNT

	T	C	A	G
1	.2071	.2929	.2071	.2929
2	.2929	.2071	.2929	.2071
3	1.	.0	.0	.0

Amino acid	Abundance	Amino acid	Abundance
A	6.06%	C	4.29% lfaa
D	8.58%	E	none
F	6.06%	G	6.06%
H	8.58%	I	6.06%
K	none	L	8.58%
M	none	N	6.06%
P	6.06%	Q	none
R	6.06%	S	8.58% mfaa
T	4.29% lfaa	V	8.58%
W	none	Y	6.06%
stop	none		

i	$(1/ratio)^i$	$(ratio)^i$	stop-free
1	2.0	.5	1.
2	4.0	.25	1.
3	8.0	.125	1.
4	16.0	.0625	1.
5	32.0	.03125	1.
6	64.0	.015625	1.
7	128.0	.0078125	1.

Table 10: Abundances obtained from various vgCodon (continued)

D. Optimized NNG

	T	C	A	G
1	.23	.21	.23	.33
2	.215	.285	.285	.215
3	.0	.0	.0	1.0

Amino acid	Abundance	Amino acid	Abundance
A	9.40%	C	none
D	none	E	9.40%
F	none	G	7.10%
H	none	I	none
K	6.60%	L	9.50% mfaa
M	4.90%	N	none
P	6.00%	Q	6.00%
R	9.50%	S	6.60%
T	6.6 %	V	7.10%
W	4.90% lfaa	Y	none
stop	6.60%		

j	$(1/ratio)^j$	$(ratio)^j$	stop-free
1	1.9388	.51579	0.934
2	3.7588	.26604	0.8723
3	7.2876	.13722	0.8148
4	14.1289	.07078	0.7610
5	27.3929	$3.65 \cdot 10^{-2}$	0.7108
6	53.109	$1.88 \cdot 10^{-2}$	0.6639
7	102.96	$9.72 \cdot 10^{-3}$	0.6200

Table 10: Abundances obtained from optimum vgCodon (continued)

E. Unoptimized NNS (NNK gives identical distribution)

	T	C	A	G
1	.25	.25	.25	.25
2	.25	.25	.25	.25
3	.0	.5	.0	0.5

Amino acid	Abundance	Amino acid	Abundance
A	6.25%	C	3.125%
D	3.125%	E	3.125%
F	3.125%	G	6.25%
H	3.125%	I	3.125%
K	3.125%	L	9.375%
M	3.125%	N	3.125%
P	6.25%	Q	3.125%
R	9.375%	S	9.375%
T	6.25%	V	6.25%
W	3.125%	Y	3.125%
stop	3.125%		

j	$(1/ratio)^j$	$(ratio)^j$	stop-free
1	3.0	.33333	.96875
2	9.0	.11111	.9385
3	27.0	.03704	.90915
4	81.0	.01234567	.8807
5	243.0	.0041152	.8532
6	729.0	$1.37 \cdot 10^{-3}$	.82655
7	2187.0	$4.57 \cdot 10^{-4}$	.8007

Table 50: Preferred IPBDs

IPBD	Number Amino Acids	Structure	Cross Links	Secreted	Source Organism	AfM
CMTI-III	26	NMR	3 SS	yes	cucumber	trypsin
ST-I_A	13	NMR	3 SS	yes	E. coli	MAbs & guanylate cyclase
α-Conotoxins	13-15	NMR	2 SS	yes	Conus snails	Receptor
μ-Conotoxins	20-25	NMR	3 SS	yes	Conus snails	Receptor
Ω-Conotoxins	25-30	-	3 SS	yes	Conus snails	Receptor
King-kong peptides	25-30	-	3 SS	yes	Conus snails	Mabs
Charybdotoxin (scorpion toxin)	37	NMR	3 SS 7-28, 13-33 17-35 (1:4, 2:5, 3:6)	yes	Leiurus quinquestriatus hebraeus	Ca^{+2} -dependent K^+ channel
Apamin (bee venom)	12	NMR	2 SS (1:3, 2:4)	yes	Bees	Mabs, Receptor(?)

Table 130: Sampling of a Library encoded by $(NNK)^6$

A. Numbers of hexapeptides in each class

total = 64,000,000 stop-free sequences.

α can be one of [WMFYCIKDENHQ]
Φ can be one of [PTAVG]
Ω can be one of [SLR]

αααααα	=	2985984.	Φααααα	=	7464960.
Ωααααα	=	4478976.	ΦΦααααα	=	7776000.
ΦΩααααα	=	9331200.	ΩΩααααα	=	2799360.
ΦΦΦααα	=	4320000.	ΦΦΩααα	=	7776000.
ΦΩΩααα	=	4665600.	ΩΩΩααα	=	933120.
ΦΦΦΦαα	=	1350000.	ΦΦΦΩαα	=	3240000.
ΦΦΩΩαα	=	2916000.	ΦΩΩΩαα	=	1166400.
ΩΩΩΩαα	=	174960.	ΦΦΦΦΦα	=	225000.
ΦΦΦΦΩα	=	675000.	ΦΦΦΩΩα	=	810000.
ΦΦΩΩΩα	=	486000.	ΦΩΩΩΩα	=	145800.
ΩΩΩΩΩα	=	17496.	ΦΦΦΦΦΦ	=	15625.
ΦΦΦΦΦΩ	=	56250.	ΦΦΦΦΩΩ	=	84375.
ΦΦΦΩΩΩ	=	67500.	ΦΦΩΩΩΩ	=	30375.
ΦΩΩΩΩΩ	=	7290.	ΩΩΩΩΩΩ	=	729.

ΦΦΩΩαα, for example, stands for the set of peptides having two amino acids from the α class, two from Φ, and two from Ω arranged in any order. There are, for example, 729 = 3^6 sequences composed entirely of S, L, and R.

Table 130: Sampling of a Library encoded by $(NNK)^6$ (continued)

B. Probability that any given stop-free DNA sequence will encode a hexapeptide from a stated class.

	P	% of class
ααααααα...	3.364E-03	(1.13E-07)
Φααααα...	1.682E-02	(2.25E-07)
Ωααααα...	1.514E-02	(3.38E-07)
ΦΦααααα...	3.505E-02	(4.51E-07)
ΦΩααααα...	6.308E-02	(6.76E-07)
ΩΩααααα...	2.839E-02	(1.01E-06)
ΦΦΦαααα...	3.894E-02	(9.01E-07)
ΦΦΩαααα...	1.051E-01	(1.35E-06)
ΦΩΩαααα...	9.463E-02	(2.03E-06)
ΩΩΩαααα...	2.839E-02	(3.04E-06)
ΦΦΦΦαααα...	2.434E-02	(1.80E-06)
ΦΦΦΩαααα...	8.762E-02	(2.70E-06)
ΦΦΩΩαααα...	1.183E-01	(4.06E-06)
ΦΩΩΩαααα...	7.097E-02	(6.08E-06)
ΩΩΩΩαααα...	1.597E-02	(9.13E-06)
ΦΦΦΦΦα...	8.113E-03	(3.61E-06)
ΦΦΦΦΩα...	3.651E-02	(5.41E-06)
ΦΦΦΩΩα...	6.571E-02	(8.11E-06)
ΦΦΩΩΩα...	5.914E-02	(1.22E-05)
ΦΩΩΩΩα...	2.661E-02	(1.83E-05)
ΩΩΩΩΩα...	4.790E-03	(2.74E-05)
ΦΦΦΦΦΦ...	1.127E-03	(7.21E-06)
ΦΦΦΦΦΩ...	6.084E-03	(1.08E-05)
ΦΦΦΦΩΩ...	1.369E-02	(1.62E-05)
ΦΦΦΩΩΩ...	1.643E-02	(2.43E-05)
ΦΦΩΩΩΩ...	1.109E-02	(3.65E-05)
ΦΩΩΩΩΩ...	3.992E-03	(5.48E-05)
ΩΩΩΩΩΩ...	5.988E-04	(8.21E-05)

Table 130: Sampling of a Library encoded by $(NNK)^6$ (continued)

C. Number of different stop-free amino-acid sequences in each class expected for various library sizes

Library size = 1.0000E+06

total = 9.7446E+05 % sampled = 1.52

Class	Number	%	Class	Number	%
αααααα...	3362.6(	.1)	Φααααα...	16803.4(	.2)
Ωααααα...	15114.6(	.3)	ΦΦααααα...	34967.8(	.4)
ΦΩααααα...	62871.1(	.7)	ΩΩαααα...	28244.3(	1.0)
ΦΦΦαααα...	38765.7(	.9)	ΦΦΩαααα...	104432.2(	1.3)
ΦΩΩαααα...	93672.7(	2.0)	ΩΩΩαααα...	27960.3(	3.0)
ΦΦΦΦααα...	24119.9(	1.8)	ΦΦΦΩααα...	86442.5(	2.7)
ΦΦΩΩααα...	115915.5(	4.0)	ΦΩΩΩααα...	68853.5(	5.9)
ΩΩΩΩααα...	15261.1(	8.7)	ΦΦΦΦΦα...	7968.1(	3.5)
ΦΦΦΦΩα...	35537.2(	5.3)	ΦΦΦΩΩα...	63117.5(	7.8)
ΦΦΩΩΩα...	55684.4(	11.5)	ΦΩΩΩΩα...	24325.9(	16.7)
ΩΩΩΩΩα...	4190.6(	24.0)	ΦΦΦΦΦΦ...	1087.1(	7.0)
ΦΦΦΦΦΩ...	5767.0(	10.3)	ΦΦΦΦΩΩ...	12637.2(	15.0)
ΦΦΦΩΩΩ...	14581.7(	21.6)	ΦΦΩΩΩΩ...	9290.2(	30.6)
ΦΩΩΩΩΩ...	3073.9(	42.2)	ΩΩΩΩΩΩ...	408.4(	56.0)

Library size = 3.0000E+06

total = 2.7885E+06 % sampled = 4.36

αααααα...	10076.4(	.3)	Φααααα...	50296.9(	.7)
Ωααααα...	45190.9(	1.0)	ΦΦααααα...	104432.2(	1.3)
ΦΩααααα...	187345.5(	2.0)	ΩΩαααα...	83880.9(	3.0)
ΦΦΦαααα...	115256.6(	2.7)	ΦΦΩαααα...	309107.9(	4.0)
ΦΩΩαααα...	275413.9(	5.9)	ΩΩΩαααα...	81392.5(	8.7)
ΦΦΦΦααα...	71074.5(	5.3)	ΦΦΦΩααα...	252470.2(	7.8)
ΦΦΩΩααα...	334106.2(	11.5)	ΦΩΩΩααα...	194606.9(	16.7)
ΩΩΩΩααα...	41905.9(	24.0)	ΦΦΦΦΦα...	23067.8(	10.3)
ΦΦΦΦΩα...	101097.3(	15.0)	ΦΦΦΩΩα...	174981.0(	21.6)
ΦΦΩΩΩα...	148643.7(	30.6)	ΦΩΩΩΩα...	61478.9(	42.2)
ΩΩΩΩΩα...	9801.0(	56.0)	ΦΦΦΦΦΦ...	3039.6(	19.5)
ΦΦΦΦΦΩ...	15587.7(	27.7)	ΦΦΦΦΩΩ...	32516.8(	38.5)
ΦΦΦΩΩΩ...	34975.6(	51.8)	ΦΦΩΩΩΩ...	20215.5(	66.6)
ΦΩΩΩΩΩ...	5879.9(	80.7)	ΩΩΩΩΩΩ...	667.0(	91.5)

Table 130: Sampling of a Library encoded by $(NNK)^6$ (continued)

Library size = 1.0000E+07

total = 8.1204E+06 % sampled = 12.69

ααααααα...	33455.9	(1.1)	Φαααααα...	166342.4 (2.2)
Ωαααααα...	148871.1	(3.3)	ΦΦαααα...	342685.7 (4.4)
ΦΩαααα...	609987.6	(6.5)	ΩΩαααα...	269958.3 (9.6)
ΦΦΦααα...	372371.8	(8.6)	ΦΦΩααα...	983416.4 (12.6)
ΦΩΩααα...	856471.6	(18.4)	ΩΩΩααα...	244761.5 (26.2)
ΦΦΦΦαα...	222702.0	(16.5)	ΦΦΦΩαα...	767692.5 (23.7)
ΦΦΩΩαα...	972324.6	(33.3)	ΦΩΩΩαα...	531651.3 (45.6)
ΩΩΩΩαα...	104722.3	(59.9)	ΦΦΦΦΦα...	68111.0 (30.3)
ΦΦΦΦΩα...	281976.3	(41.8)	ΦΦΦΩΩα...	450120.2 (55.6)
ΦΦΩΩΩα...	342072.1	(70.4)	ΦΩΩΩΩα...	122302.6 (83.9)
ΩΩΩΩΩα...	16364.0	(93.5)	ΦΦΦΦΦΦ...	8028.0 (51.4)
ΦΦΦΦΦΩ...	37179.9	(66.1)	ΦΦΦΦΩΩ...	67719.5 (80.3)
ΦΦΦΩΩΩ...	61580.0	(91.2)	ΦΦΩΩΩΩ...	29586.1 (97.4)
ΦΩΩΩΩΩ...	7259.5	(99.6)	ΩΩΩΩΩΩ...	728.8 (100.0)

Library size = 3.0000E+07

total = 1.8633E+07 % sampled = 29.11

ααααααα...	99247.4	(3.3)	Φαααααα...	487990.0 (6.5)
Ωαααααα...	431933.3	(9.6)	ΦΦαααα...	983416.5 (12.6)
ΦΩαααα...	1712943.0	(18.4)	ΩΩαααα...	734284.6 (26.2)
ΦΦΦααα...	1023590.0	(23.7)	ΦΦΩααα...	2592866.0 (33.3)
ΦΩΩααα...	2126605.0	(45.6)	ΩΩΩααα...	558519.0 (59.9)
ΦΦΦΦαα...	563952.6	(41.8)	ΦΦΦΩαα...	1800481.0 (55.6)
ΦΦΩΩαα...	2052433.0	(70.4)	ΦΩΩΩαα...	978420.5 (83.9)
ΩΩΩΩαα...	163640.3	(93.5)	ΦΦΦΦΦα...	148719.7 (66.1)
ΦΦΦΦΩα...	541755.7	(80.3)	ΦΦΦΩΩα...	738960.1 (91.2)
ΦΦΩΩΩα...	473377.0	(97.4)	ΦΩΩΩΩα...	145189.7 (99.6)
ΩΩΩΩΩα...	17491.3	(100.0)	ΦΦΦΦΦΦ...	13829.1 (88.5)
ΦΦΦΦΦΩ...	54058.1	(96.1)	ΦΦΦΦΩΩ...	83726.0 (99.2)
ΦΦΦΩΩΩ...	67454.5	(99.9)	ΦΦΩΩΩΩ...	30374.5 (100.0)
ΦΩΩΩΩΩ...	7290.0	(100.0)	ΩΩΩΩΩΩ...	729.0 (100.0)

Table 130: Sampling of a Library encoded by $(NNK)^6$ (continued)

Library size = 7.6000E+07

total = 3.2125E+07 % sampled = 50.19

ααααααα...	245057.8(8.2)	Φαααααα...	1175010.0(15.7)
Ωαααααα...	1014733.0(22.7)	ΦΦαααααα...	2255280.0(29.0)
ΦΩαααααα...	3749112.0(40.2)	ΩΩαααααα...	1504128.0(53.7)
ΦΦΦαααα...	2142478.0(49.6)	ΦΦΩαααα...	4993247.0(64.2)
ΦΩΩαααα...	3666785.0(78.6)	ΩΩΩαααα...	840691.9(90.1)
ΦΦΦΦααα...	1007002.0(74.6)	ΦΦΦΩααα...	2825063.0(87.2)
ΦΦΩΩααα...	2782358.0(95.4)	ΦΩΩΩααα...	1154956.0(99.0)
ΩΩΩΩααα...	174790.0(99.9)	ΦΦΦΦΦα...	210475.6(93.5)
ΦΦΦΦΩα...	663929.3(98.4)	ΦΦΦΩΩα...	808298.6(99.8)
ΦΦΩΩΩα...	485953.2(100.0)	ΦΩΩΩΩα...	145799.9(100.0)
ΩΩΩΩΩα...	17496.0(100.0)	ΦΦΦΦΦΦ...	15559.9(99.6)
ΦΦΦΦΦΩ...	56234.9(100.0)	ΦΦΦΦΩΩ...	84374.6(100.0)
ΦΦΦΩΩΩ...	67500.0(100.0)	ΦΦΩΩΩΩ...	30375.0(100.0)
ΦΩΩΩΩΩ...	7290.0(100.0)	ΩΩΩΩΩΩ...	729.0(100.0)

Library size = 1.0000E+08

total = 3.6537E+07 % sampled = 57.09

ααααααα...	318185.1(10.7)	Φαααααα...	1506161.0(20.2)
Ωαααααα...	1284677.0(28.7)	ΦΦαααααα...	2821285.0(36.3)
ΦΩαααααα...	4585163.0(49.1)	ΩΩαααααα...	1783932.0(63.7)
ΦΦΦαααα...	2566085.0(59.4)	ΦΦΩαααα...	5764391.0(74.1)
ΦΩΩαααα...	4051713.0(86.8)	ΩΩΩαααα...	888584.3(95.2)
ΦΦΦΦααα...	1127473.0(83.5)	ΦΦΦΩααα...	3023170.0(93.3)
ΦΦΩΩααα...	2865517.0(98.3)	ΦΩΩΩααα...	1163743.0(99.8)
ΩΩΩΩααα...	174941.0(100.0)	ΦΦΦΦΦα...	218886.6(97.3)
ΦΦΦΦΩα...	671976.9(99.6)	ΦΦΦΩΩα...	809757.3(100.0)
ΦΦΩΩΩα...	485997.5(100.0)	ΦΩΩΩΩα...	145800.0(100.0)
ΩΩΩΩΩα...	17496.0(100.0)	ΦΦΦΦΦΦ...	15613.5(99.9)
ΦΦΦΦΦΩ...	56248.9(100.0)	ΦΦΦΦΩΩ...	84375.0(100.0)
ΦΦΦΩΩΩ...	67500.0(100.0)	ΦΦΩΩΩΩ...	30375.0(100.0)
ΦΩΩΩΩΩ...	7290.0(100.0)	ΩΩΩΩΩΩ...	729.0(100.0)

Table 130: Sampling of a Library encoded by $(NNK)^6$ (continued)

Library size = 3.0000E+08

total = 5.2634E+07 % sampled = 82.24

ααααααα...	856451.3(28.7)	Φαααααα...	3668130.0(49.1)
Ωαααααα...	2854291.0(63.7)	ΦΦαααααα...	5764391.0(74.1)
ΦΩαααααα...	8103426.0(86.8)	ΩΩαααααα...	2665753.0(95.2)
ΦΦΦαααα...	4030893.0(93.3)	ΦΦΩαααα...	7641378.0(98.3)
ΦΩΩαααα...	4654972.0(99.8)	ΩΩΩαααα...	933018.6(100.0)
ΦΦΦΦαααα...	1343954.0(99.6)	ΦΦΦΩαααα...	3239029.0(100.0)
ΦΦΩΩαααα...	2915985.0(100.0)	ΦΩΩΩαααα...	1166400.0(100.0)
ΩΩΩΩαααα...	174960.0(100.0)	ΦΦΦΦΦα...	224995.5(100.0)
ΦΦΦΦΩα...	674999.9(100.0)	ΦΦΦΩΩα...	810000.0(100.0)
ΦΦΩΩΩα...	486000.0(100.0)	ΦΩΩΩΩα...	145800.0(100.0)
ΩΩΩΩΩα...	17496.0(100.0)	ΦΦΦΦΦΦ...	15625.0(100.0)
ΦΦΦΦΦΩ...	56250.0(100.0)	ΦΦΦΦΩΩ...	84375.0(100.0)
ΦΦΦΩΩΩ...	67500.0(100.0)	ΦΦΩΩΩΩ...	30375.0(100.0)
ΦΩΩΩΩΩ...	7290.0(100.0)	ΩΩΩΩΩΩ...	729.0(100.0)

Library size = 1.0000E+09

total = 6.1999E+07 % sampled = 96.87

ααααααα...	2018278.0(67.6)	Φαααααα...	6680917.0(89.5)
Ωαααααα...	4326519.0(96.6)	ΦΦαααααα...	7690221.0(98.9)
ΦΩαααααα...	9320389.0(99.9)	ΩΩαααααα...	2799250.0(100.0)
ΦΦΦαααα...	4319475.0(100.0)	ΦΦΩαααα...	7775990.0(100.0)
ΦΩΩαααα...	4665600.0(100.0)	ΩΩΩαααα...	933120.0(100.0)
ΦΦΦΦαααα...	1350000.0(100.0)	ΦΦΦΩαααα...	3240000.0(100.0)
ΦΦΩΩαααα...	2916000.0(100.0)	ΦΩΩΩαααα...	1166400.0(100.0)
ΩΩΩΩαααα...	174960.0(100.0)	ΦΦΦΦΦα...	225000.0(100.0)
ΦΦΦΦΩα...	675000.0(100.0)	ΦΦΦΩΩα...	810000.0(100.0)
ΦΦΩΩΩα...	486000.0(100.0)	ΦΩΩΩΩα...	145800.0(100.0)
ΩΩΩΩΩα...	17496.0(100.0)	ΦΦΦΦΦΦ...	15625.0(100.0)
ΦΦΦΦΦΩ...	56250.0(100.0)	ΦΦΦΦΩΩ...	84375.0(100.0)
ΦΦΦΩΩΩ...	67500.0(100.0)	ΦΦΩΩΩΩ...	30375.0(100.0)
ΦΩΩΩΩΩ...	7290.0(100.0)	ΩΩΩΩΩΩ...	729.0(100.0)

Table 130: Sampling of a Library encoded by $(NNK)^6$ (continued)

Library size = 3.0000E+09

total = 6.3890E+07 % sampled = 99.83

αααααα...	2884346.0	(96.6)	Φααααα...	7456311.0 (99.9)
Ωααααα...	4478800.0	(100.0)	ΦΦαααα...	7775990.0 (100.0)
ΦΩααααα...	9331200.0	(100.0)	ΩΩαααα...	2799360.0 (100.0)
ΦΦΦαααα...	4320000.0	(100.0)	ΦΦΩααα...	7776000.0 (100.0)
ΦΩΩαααα...	4665600.0	(100.0)	ΩΩΩααα...	933120.0 (100.0)
ΦΦΦΦαα...	1350000.0	(100.0)	ΦΦΦΩαα...	3240000.0 (100.0)
ΦΦΩΩαα...	2916000.0	(100.0)	ΦΩΩΩαα...	1166400.0 (100.0)
ΩΩΩΩαα...	174960.0	(100.0)	ΦΦΦΦΦα...	225000.0 (100.0)
ΦΦΦΦΩα...	675000.0	(100.0)	ΦΦΦΩΩα...	810000.0 (100.0)
ΦΦΩΩΩα...	486000.0	(100.0)	ΦΩΩΩΩα...	145800.0 (100.0)
ΩΩΩΩΩα...	17496.0	(100.0)	ΦΦΦΦΦΦ...	15625.0 (100.0)
ΦΦΦΦΦΩ...	56250.0	(100.0)	ΦΦΦΦΩΩ...	84375.0 (100.0)
ΦΦΦΩΩΩ...	67500.0	(100.0)	ΦΦΩΩΩΩ...	30375.0 (100.0)
ΦΩΩΩΩΩ...	7290.0	(100.0)	ΩΩΩΩΩΩ...	729.0 (100.0)

Table 130, continued

D. Formulae for tabulated quantities.

Lsize is the number of independent transformants.
31**6 is 31 to sixth power; 6*3 means 6 times 3.
A = Lsize/(31**6)
α can be one of [WMFYCIKDENHQ.]
Φ can be one of [PTAVG]
Ω can be one of [SLR]

F0 = (12)**6 F1 = (12)**5 F2 = (12)**4
F3 = (12)**3 F4 = (12)**2 F5 = (12)
F6 = 1

ααααααα = F0 * (1-exp(-A))
Φαααααα = 6 * 5 * F1 * (1-exp(-2*A))
Ωαααααα = 6 * 3 * F1 * (1-exp(-3*A))
ΦΦαααααα = (15) * 5**2 * F2 * (1-exp(-4*A))
ΦΩαααααα = (6*5)*5*3 *F2 * (1-exp(-6*A))
ΩΩαααααα = (15) * 3**2 * F2 * (1-exp(-9*A))
ΦΦΦαααα = (20)*(5**3) * F3 * (1-exp(-8*A))
ΦΦΩαααα = (60)*(5*5*3)*F3* (1-exp(-12*A))
ΦΩΩαααα = (60)*(5*3*3)*F3*(1-exp(-18*A))
ΩΩΩαααα = (20)*(3)**3*F3*(1-exp(-27*A))
ΦΦΦΦααα = (15)*(5)**4*F4*(1-exp(-16*A))
ΦΦΦΩααα = (60)*(5)**3*3*F4*(1-exp(-24*A))
ΦΦΩΩααα = (90)*(5*5*3*3)*F4*(1-exp(-36*A))
ΦΩΩΩααα = (60)*(5*3*3*3)*F4*(1-exp(-54*A))
ΩΩΩΩααα = (15)*(3)**4 * F4 *(1-exp(-81*A))
ΦΦΦΦΦαα = (6)*(5)**5 * F5 * (1-exp(-32*A))
ΦΦΦΦΩαα = 30*5*5*5*5*3*F5*(1-exp(-48*A))
ΦΦΦΩΩαα = 60*5*5*5*3*3*F5*(1-exp(-72*A))
ΦΦΩΩΩαα = 60*5*5*3*3*3*F5*(1-exp(-108*A))
ΦΩΩΩΩαα = 30*5*3*3*3*3*F5*(1-exp(-162*A))
ΩΩΩΩΩαα = 6*3*3*3*3*3*F5*(1-exp(-243*A))
ΦΦΦΦΦΦ = 5**6 * (1-exp(-64*A))
ΦΦΦΦΦΩ = 6*3*5**5*(1-exp(-96*A))
ΦΦΦΦΩΩ = 15*3*3*5**4*(1-exp(-144*A))
ΦΦΦΩΩΩ = 20*3**3*5**3*(1-exp(-216*A))
ΦΦΩΩΩΩ = 15*3**4*5**2*(1-exp(-324*A))
ΦΩΩΩΩΩ = 6*3**5*5*(1-exp(-486*A))
ΩΩΩΩΩΩ = 3**6*(1-exp(-729*A))
total = ααααααα + Φαααααα + Ωαααααα + ΦΦαααααα + ΦΩαααααα +
ΩΩαααααα + ΦΦΦαααα + ΦΦΩαααα + ΦΩΩαααα + ΩΩΩαααα +
ΦΦΦΦααα + ΦΦΦΩααα + ΦΦΩΩααα + ΦΩΩΩααα + ΩΩΩΩααα +
ΦΦΦΦΦαα + ΦΦΦΦΩαα + ΦΦΦΩΩαα + ΦΦΩΩΩαα + ΦΩΩΩΩαα +
ΩΩΩΩΩαα + ΦΦΦΦΦΦ + ΦΦΦΦΦΩ + ΦΦΦΦΩΩ + ΦΦΦΩΩΩ +
ΦΦΩΩΩΩ + ΦΩΩΩΩΩ + ΩΩΩΩΩΩ

Table 131: Sampling of a Library Encoded by $(NNT)^4(NNG)^2$

X can be F,S,Y,C,L,P,H,R,I,T,N,V,A,D,G

Γ can be L^2,R^2,S,W,P,Q,M,T,K,V,A,E,G

Library comprises $8.55 \cdot 10^6$ amino-acid sequences; $1.47 \cdot 10^7$ DNA sequences.

Total number of possible aa sequences= 8,555,625

x	LVPTARGFYCHIND
S	S
θ	VPTAGWQMKES
Ω	LR

The first, second, fifth, and sixth positions can hold x or S; the third and fourth position can hold θ or Ω. I have lumped sequences by the number of xs, Ss, θs, and Ωs.

For example xxθΩSS stands for:
[xxθΩSS, xSθΩxS, xSθΩSx, SSθΩxx, SxθΩxS, SxθΩSx,
xxΩθSS, xSΩθxS, xSΩθSx, SSΩθxx, SxΩθxS, SxΩθSx]

The following table shows the likelihood that any particular DNA sequence will fall into one of the defined classes.

Library size = 1.0 Sampling = .00001%

total.....	1.0000E+00	%sampled.....	1.1688E-07
xxθθxx.....	3.1524E-01	xxθΩxx.....	2.2926E-01
xxΩΩxx.....	4.1684E-02	xxθθxS.....	1.8013E-01
xxθΩxS.....	1.3101E-01	xxΩΩxS.....	2.3819E-02
xxθθSS.....	3.8600E-02	xxθΩSS.....	2.8073E-02
xxΩΩSS.....	5.1042E-03	xSθθSS.....	3.6762E-03
xSθΩSS.....	2.6736E-03	xSΩΩSS.....	4.8611E-04
SSθθSS.....	1.3129E-04	SSθΩSS.....	9.5486E-05
SSΩΩSS.....	1.7361E-05		

Table 131: Sampling of a Library
Encoded by $(NNT)^4(NNG)^2$
(continued)

The following sections show how many sequences of each class are expected for libraries of different sizes.

Library size = 1.0000E+05

total..... 9.9137E+04 fraction sampled = 1.1587E-02

Type	Number	%	Type	Number	%
xx0Ωxx.....	22771.4	(1.3)	xx00xx.....	31416.9	(.7)
xx0Ωxs.....	17891.8	(1.3)	xxΩΩxx.....	4112.4	(2.7)
xxΩΩxs.....	2318.5	(5.3)	xx0Ωxs.....	12924.6	(2.7)
xx0ΩSS.....	2732.5	(5.3)	xx00SS.....	3808.1	(2.7)
xS00SS.....	357.8	(5.3)	xxΩΩSS.....	483.7	(10.3)
xSΩΩSS.....	43.7	(19.5)	xS0ΩSS.....	253.4	(10.3)
SS0ΩSS.....	8.6	(19.5)	SS00SS.....	12.4	(10.3)
			SSΩΩSS.....	1.4	(35.2)

Library size = 1.0000E+06

total..... 9.2064E+05 fraction sampled = 1.0761E-01

Type	Number	%	Type	Number	%
xx0Ωxx.....	214394.0	(12.7)	xx00xx.....	304783.9	(6.6)
xx0Ωxs.....	168452.5	(12.7)	xxΩΩxx.....	36508.6	(23.8)
xx0ΩSS.....	21666.6	(41.9)	xx0Ωxs.....	114741.4	(23.8)
xxΩΩxs.....	18383.8	(41.9)	xx00SS.....	33807.7	(23.8)
xS00SS.....	2837.3	(41.9)	xxΩΩSS.....	3114.6	(66.2)
xSΩΩSS.....	198.4	(88.6)	xS0ΩSS.....	1631.5	(66.2)
SS0ΩSS.....	39.0	(88.6)	SS00SS.....	80.1	(66.2)
			SSΩΩSS.....	3.9	(98.7)

Library size = 3.0000E+06

total..... 2.3880E+06 fraction sampled = 2.7912E-01

Type	Number	%	Type	Number	%
xx0Ωxx.....	565051.6	(33.4)	xx00xx.....	855709.5	(18.4)
xx0Ωxs.....	443969.1	(33.4)	xxΩΩxx.....	85564.7	(55.7)
xx0ΩSS.....	41581.5	(80.4)	xx0Ωxs.....	268917.8	(55.7)
xxΩΩxs.....	35281.3	(80.4)	xx00SS.....	79234.7	(55.7)
xS00SS.....	5445.2	(80.4)	xxΩΩSS.....	4522.6	(96.1)
xSΩΩSS.....	223.7	(99.9)	xS0ΩSS.....	2369.0	(96.1)
SS0ΩSS.....	43.9	(99.9)	SS00SS.....	116.3	(96.1)
			SSΩΩSS.....	4.0	(100.0)

Table 131: Sampling of a Library
Encoded by $(NNT)^4(NNG)^2$
(continued)

Library size = 8.5556E+06

total.....	4.9303E+06	fraction sampled = 5.7626E-01	
xxθθxx.....	2046301.0(44.0)	xxθΩxx.....	1160645.0(68.7)
xxΩΩxx.....	138575.9(90.2)	xxθθxS.....	911935.6(68.7)
xxθΩxS.....	435524.3(90.2)	xxΩΩxS.....	43480.7(99.0)
xxθθSS.....	128324.1(90.2)	xxθΩSS.....	51245.1(99.0)
xxΩΩSS.....	4703.6(100.0)	xSθθSS.....	6710.7(99.0)
xSθΩSS.....	2463.8(100.0)	xSΩΩSS.....	224.0(100.0)
SSθθSS.....	121.0(100.0)	SSθΩSS.....	44.0(100.0)
SSΩΩSS.....	4.0(100.0)		

Library size = 1.0000E+07

total.....	5.3667E+06	fraction sampled = 6.2727E-01	
xxθθxx.....	2289093.0(49.2)	xxθΩxx.....	1254877.0(74.2)
xxΩΩxx.....	143467.0(93.4)	xxθθxS.....	985974.9(74.2)
xxθΩxS.....	450896.3(93.4)	xxΩΩxS.....	43710.7(99.6)
xxθθSS.....	132853.4(93.4)	xxθΩSS.....	51516.1(99.6)
xxΩΩSS.....	4703.9(100.0)	xSθθSS.....	6746.2(99.6)
xSθΩSS.....	2464.0(100.0)	xSΩΩSS.....	224.0(100.0)
SSθθSS.....	121.0(100.0)	SSθΩSS.....	44.0(100.0)
SSΩΩSS.....	4.0(100.0)		

Library size = 3.0000E+07

total.....	7.8961E+06	fraction sampled = 9.2291E-01	
xxθθxx.....	4040589.0(86.9)	xxθΩxx.....	1661409.0(98.3)
xxΩΩxx.....	153619.1(100.0)	xxθθxS.....	1305393.0(98.3)
xxθΩxS.....	482802.9(100.0)	xxΩΩxS.....	43904.0(100.0)
xxθθSS.....	142254.4(100.0)	xxθΩSS.....	51744.0(100.0)
xxΩΩSS.....	4704.0(100.0)	xSθθSS.....	6776.0(100.0)
xSθΩSS.....	2464.0(100.0)	xSΩΩSS.....	224.0(100.0)
SSθθSS.....	121.0(100.0)	SSθΩSS.....	44.0(100.0)
SSΩΩSS.....	4.0(100.0)		

Table 131: Sampling of a Library
Encoded by $(NNT)^4(NNG)^2$
(continued)

Library size = 5.0000E+07

total.....	8.3956E+06	fraction sampled = 9.8130E-01		
xx0θxx.....	4491779.0(96.6)	xx0Ωxx.....	1688387.0(99.9)	
xxΩΩxx.....	153663.8(100.0)	xx0θxS.....	1326590.0(99.9)	
xx0ΩxS.....	482943.4(100.0)	xxΩΩxS.....	43904.0(100.0)	
xx0θSS.....	142295.8(100.0)	xx0ΩSS.....	51744.0(100.0)	
xxΩΩSS.....	4704.0(100.0)	xS0θSS.....	6776.0(100.0)	
xS0ΩSS.....	2464.0(100.0)	xSΩΩSS.....	224.0(100.0)	
SS0θSS.....	121.0(100.0)	SS0ΩSS.....	44.0(100.0)	
SSΩΩSS.....	4.0(100.0)			

Library size = 1.0000E+08

total.....	8.5503E+06	fraction sampled = 9.9938E-01		
xx0θxx.....	4643063.0(99.9)	xx0Ωxx.....	1690302.0(100.0)	
xxΩΩxx.....	153664.0(100.0)	xx0θxS.....	1328094.0(100.0)	
xx0ΩxS.....	482944.0(100.0)	xxΩΩxS.....	43904.0(100.0)	
xx0θSS.....	142296.0(100.0)	xx0ΩSS.....	51744.0(100.0)	
xxΩΩSS.....	4704.0(100.0)	xS0θSS.....	6776.0(100.0)	
xS0ΩSS.....	2464.0(100.0)	xSΩΩSS.....	224.0(100.0)	
SS0θSS.....	121.0(100.0)	SS0ΩSS.....	44.0(100.0)	
SSΩΩSS.....	4.0(100.0)			

Table 132: Relative efficiencies of various simple variegation codons

	Number of codons		
	5	6	7
vgCodon	#DNA/#AA [#DNA] (#AA)	#DNA/#AA [#DNA] (#AA)	#DNA/#AA [#DNA] (#AA)
NNK	8.95	13.86	21.49
assuming	$[2.86 \cdot 10^7]$	$[8.87 \cdot 10^8]$	$[2.75 \cdot 10^{10}]$
stops vanish	$(3.2 \cdot 10^6)$	$(6.4 \cdot 10^7)$	$(1.28 \cdot 10^9)$
NNT	1.38	1.47	1.57
	$[1.05 \cdot 106]$	$[1.68 \cdot 10^7]$	$[2.68 \cdot 10^8]$
	$(7.59 \cdot 10^5)$	$(1.14 \cdot 10^7)$	$(1.71 \cdot 10^8)$
NNG	2.04	2.36	2.72
assuming	$[7.59 \cdot 10^5]$	$[1.14 \cdot 10^6]$	$[1.71 \cdot 10^8]$
stops vanish	$(3.7 \cdot 10^5)$	$(4.83 \cdot 10^6)$	$(6.27 \cdot 10^7)$

Table 155
Distance in Å between alpha carbons in octapeptides:

Extended Strand: angle of $C_\alpha 1$-$C_\alpha 2$-$C_\alpha 3$ = 138°

	1	2	3	4	5	6	7	8
1	-							
2		3.8	-					
3		7.1	3.8	-				
4	10.7	7.1	3.8	-				
5	14.2	10.7	7.1	3.8	-			
6	17.7	14.1	10.7	7.1	3.8	-		
7	21.2	17.7	14.1	10.6	7.0	3.8	-	
8	24.6	20.9	17.5	13.9	10.6	7.0	3.8	-

Reverse turn between residues 4 and 5.

	1	2	3	4	5	6	7	8
1	-							
2		3.8	-					
3		7.1	3.8	-				
4	10.6	7.0	3.8	-				
5	11.6	8.0	6.1	3.8	-			
6	9.0	5.8	5.5	5.6	3.8	-		
7	6.2	4.1	6.3	8.0	7.0	3.8	-	
8	5.8	6.0	9.1	11.6	10.7	7.2	3.8	-

Alpha helix: angle of $C_\alpha 1$-$C_\alpha 2$-$C_\alpha 3$ = 93°

	1	2	3	4	5	6	7	8
1	-							
2		3.8	-					
3		5.5	3.8	-				
4	5.1	5.4	3.8	-				
5	6.6	5.3	5.5	3.8	-			
6	9.3	7.0	5.6	5.5	3.8	-		
7	10.4	9.3	6.9	5.4	5.5	3.8	-	
8	11.3	10.7	9.5	6.8	5.6	5.6	3.8	-

Table 156

Distances between alpha carbons in closed mini-proteins of the form disulfide cyclo(CXXXXC)

Minimum distance

	1	2	3	4	5	6
1	-					
2	3.8	-				
3	5.9	3.8	-			
4	5.6	6.0	3.8	-		
5	4.7	5.9	6.0	3.8	-	
6	4.8	5.3	5.1	5.2	3.8	-

Average distance

	1	2	3	4	5	6
1	-					
2	3.8	-				
3	6.3	3.8	-			
4	7.5	6.4	3.8	-		
5	7.1	7.5	6.3	3.8	-	
6	5.6	7.5	7.7	6.4	3.8	-

Maximum distance

	1	2	3	4	5	6
1	-					
2	3.8	-				
3	6.7	3.8	-			
4	9.0	6.9	3.8	-		
5	8.7	8.8	6.8	3.8	-	
6	6.6	9.2	9.1	6.8	3.8	-

Table 820: Peptide Phage

Name	Putative Streptavidin Binding Peptide Seq.	Antibiotic Resistance Marker
HPQ	A E G P C H P Q F - - C Q S Y I E G R I V - - - - - E...	
DEV(F)	A E - P C H P Q Y R L C Q R P L K Q P P P P P P P A E...	
Dev(E)	A E - L C H P Q F P R C N L F R K V P P P P P P P A E...	
HPQ6	A E G P C H P Q F P R C Y I E G R I V - - - - - - - E...	

											1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2
1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	
-	-	-	-	C	-	-	-	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	-	E	

Table 838: Streptavidin-binding disulfide-constrained peptides

clone	glu	gly	V	cys	V	V	V	V	cys	V	ser	Frequency
#2	glu	gly	tyr	cys	his	pro	gln	phe	cys	pro	ser	4
#4	glu	gly	his	cys	his	pro	gln	phe	cys	ser	ser	3
#5	glu	gly	leu	cys	his	pro	gln	phe	cys	gly	ser	2
#8	glu	gly	asp	cys	his	pro	gln	phe	cys	ser	ser	2
#1	glu	gly	asn	cys	his	pro	gln	phe	cys	pro	ser	1
#3	glu	gly	asp	cys	his	pro	gln	phe	cys	arg	ser	1
#13	glu	gly	asp	cys	his	pro	gln	phe	cys	val	ser	1
				cys	his	pro	gln	phe	cys			consensus

Table 839: Sequences Obtained by Enrichment over BSA

clone	glu	gly	V	cys	V	V	V	V	cys	V	ser	Frequency
#21	glu	gly	gly	cys	phe	lys	arg	asn	cys	tyr	ser	1
#22	glu	gly	his	cys	asp	lys	lys	ile	cys	leu	ser	1
#23	glu	gly	phe	cys	his	thr	ala	ala	cys	phe	ser	1
#24	glu	gly	his	cys	tyr	lys	gly	val	cys	ser	ser	1
#25	glu	gly	his	cys	asp	lys	trp	arg	cys	pro	ser	1
#26	glu	gly	ile	cys	tyr	arg	leu	asp	cys	ile	ser	1
#27	glu	gly	gly	cys	phe	pro	trp	his	cys	phe	ser	1
#28	glu	gly	ser	cys	asp	ser	leu	arg	cys	asp	ser	1

No consensus observed.

CITATIONS

ALMA83a:
Almassy, RC, JC Fontecilla-Camps, FL Suddath, and CE Bugg,
Entry 1SN3 in Brookhaven Protein Data Bank, (1983).

ALMA83b:
Almassy, RC, JC Fontecilla-Camps, FL Suddath, and CE Bugg,
J Mol Biol (1983), 170:497ff.

ALMQ89:
Almquist, RG, SR Kadambi, DM Yasuda, FL Weitl, WE Polgar, and LR Toll,
Int J Pept Protein Res, (Dec 1989), 34(6)455-62.

BAKE87:
Baker, K, N Mackman, and IB Holland,
Prog Biophys molec Biol (1987), 49:89-115.

BASS90:
Bass, Greene, and Wells,
Proteins: Structure, Function and Genetics (1990) 8:309-14.

BECK80:
Beck, E,
Nucl Acid Res (1980), 8(13)3011-3024.

BECK83:
Beckwith, J, and TJ Silhavy,
Methods in Enzymology (1983), 97:3-11.

BECK89b:
Becker, S, E Atherton, H Michel, and RD Gordon,
J Protein Chem, (Jun 1989), 8(3)393-4.

BECK89c:
Becker, S, E Atherton, and RD Gordon,
Eur J Biochem, (Oct 20 1989), 185(1)79-84.

BENS84:
Benson, SA, E Bremer, and TJ Silhavy,
Proc Natl Acad Sci USA (1984), 81:3830-34.

BENS88:
Benson, SA, JL Occi, BA Sampson,
J Mol Biol (1988) 203(4)961-70.

BENZ88a:
Benz, R, and K Bauer,
Eur J Biochem (1988), 176:1-19.

BENZ88b:
Benz, R,
Ann Rev Microbiol (1988), 42:359-93.

BERG88:
Berg, JM,
Proc Natl Acad Sci USA (1988), 85:99-102.

BETT88:
Better, M, CP Chang, RR Robinson, and AH Horwitz,
Science (1988), 240:1041-1043.

BHAT86:
Bhatnagar, PK, and JC Frantz,
Develop biol Standard (1986), 63:79-87.

BODE89:
Bode, W, HJ Greyling, R Huber, J Otlewski, and T Wilusz,
FEBS Lett (Jan 2 1989), 242(2)285-92.

BOEK80: {restore citation; effects of mutations near SP-1 cleavage site.}
Boeke, JD, M Russel, and P Model,
J Mol Biol (1980), 144:103-116.

BOEK82: {restore citation; effect of gIIIp on membrane}
Boeke, JD, P Model, and ND Zinder,
Molec and Gen Genet, (1982), 186:185-192.

CARM90:
Carmel, G, D Hellstern, D Henning, and JW Coulton,
J Bacteriol, (Apr 1990), 172(4)1861-9.

CARU85:
Caruthers, MH,
Science (1985), 230:281-285.

CARU87:
Caruthers, MH, P Gottlieb, LP Bracco, and L Cummings,
in Protein Structure, Folding, and Design 2, 1987,
Ed. D Oxender (New York, AR Liss Inc) p.9ff.

CATR87:
Catron, KM, and CA Schnaitman,
J Bacteriol (1987), 169:4327-34.

CHAR86a:
Charbit, A, JC Boulain, A Ryter, and M Hofnung,
EMBO J, (1986), 5(11)3029-37.

CHAR86b:
Charbit, A, J-C Boulain, and M Hofnung,
Comptes Rendu Acad Sci, Paris, (1986), 302:617-24.

CHAR87:
Charbit, A, E Sobczak, ML Michel, A Molla, P Tiollais, and M Hofnung,
J Immunol (1987), 139:1658-64.

CHAR88b:
Charbit, A, A Molla, W Saurin, and M Hofnung,
Gene (1988), 70(1)181-9.

CHAR88c:
Charbit, A, S Van der Werf, V Mimic, JC Boulain, M Girard, and M Hofnung,
Ann Inst Pasteur Microbiol (1988), 139(1)45-58.

CHAV88:
Chavrier, P, P Lemaire, O Revelant, R Bravo, and P Charnay,
Molec Cell Biol (1988), 8(3)1319-26.

CHOU74:
Chou, PY, and GD Fasman,
Biochemistry (1974), 13:(2)222-45.

CHOW87:
Chowdhuury, K, U Deutsch, and P Gruss,
Cell (1987), 48:771-778.

CLEM81:
Clement, JM, and M Hofnung,
Cell (1981), 27:507-514.

CLIC88:
Click, EM, GA McDonald, and CA Schnaitman,
J Bacteriol (1988), 170:2005-2011.

CLUN84:
Clune, A, K-S Lee, and T Ferenci,
Biochem and Biophys Res Comm (1984), 121:34-40.

CREI84:
Creighton, TE,
Proteins: Structures and Molecular Principles,
W H Freeman & Co, New York, 1984.

CREI88:
Creighton, TE,
BioEssays (1988), 8(2)57-63.

CRUZ85:
Cruz, LJ, WR Gray, BM Olivera, RD Zeikus, L Kerr, D Yoshikami, and E Moczydlowski,
J Biol Chem, (1985), 260(16)9280-8.

CRUZ89:
Cruz, LJ, G Kupryszewski, GW LeCheminant, WR Grey, BM Oliveria, and J Rivier,
Biochem (1989), 28:3437-3442.

CWIR90:
Cwirla, SE, EA Peters, RW Barrett, and WJ Dower,
Proc Natl Acad Sci USA, (August 1990), 87:6378-6382.

DALL90:
Dallas, WS,
J Bacteriol (1990), 172(9)5490-93.

DEGE84:
de Geus, P, HM Verheij, NH Reigman, WPM Hoekstra, and GH de Haas,
EMBO J (1984), 3(8)1799-1802.

DELA88:
de la Cruz, VF, AA Lal and TF McCutchan,
J Biol Chem, (1988), 263(9)4318-22.

DEVL90:
Devlin, JJ, LC Panganiban, and PE Devlin,
Science, (27 July 1990), 249:404-406.

DILL87:
Dill, KA,
Protein Engineering (1987), 1:369-371.

DUCH88:
Duchene, M, A Schweized, F Lottspeich, G Krauss, M Marget, K Vogel, B-U von Specht, and H Domdey,
J Bacteriol (1987), 170:155-162.

DULB86:
Dulbecco, R,
US Patent 4,593,002, June 3, 1986.

DWAR89:
Dwarakanath, P, SS Viswiswariah, YVBK Subrahmanyam, G Shanthi, HM Jagannatha, and TS Balganesh,
Gene (1989), 81:219-226.

EISE85:
Eisenbeis, SJ, MS Nasoff, SA Noble, LP Bracco, DR Dodds, MH Caruthers,
Proc Natl Acad Sci USA (1985), 82:1084-1088.

ELLE88:
Elleman, TC,
Microbiol Rev (1988), 52(2)233-47.

EVAN88:
Evans, RM, and SM Hollenberg,
Cell (1988), 52:1-3.

FAVE89:
Favel, A, D Le-Nguyen, MA Coletti-Previero, and C Castro,
Biochem Biophys Res Comm (1989), 162:79-82.

FERE82a:
Ferenci, T,
Ann Microbiol (Inst Pasteur) (1982), 133A:167-169.

FERE82b:
Ferenci, T, and K-S Lee,
J Mol Biol (1982), 160:431-444.

FERE83:
Ferenci, T, and KS Lee,
J Bacteriol (1983), 154:984-987.

FERE84:
Ferenci, T,
Trends in Biological Science (1984) Vol. ?:44-48.

FRAN87:
Frankel, AD, JM Berg, and CO Pabo,
Proc Natl Acad Sci USA (1987), 84:4841-45.

FREU89:
Freudl, R, H Schwarz, M Degen, and U Henning,
J Mol Biol (1989), 205(4)771-5.

GARI87:
Gariepy, J, AK Judd, and GK Schoolnik,
Proc Natl Acad Sci USA (1987), 84:8907-11.

GAUS87:
Gauss, P, KB Krassa, DS McPheeters, MA Nelson, and L Gold,
Proc Natl Acad Sci USA (1987), 84:8515-19.

GETZ88:
Getzoff, ED, HE Parge, DE McRee, and JA Tainer,
Rev Infect Dis (1988), 10(Suppl 2)S296-299.

GIBS88:
Gibson, TJ, JPM Postma, RS Brown, and P Argos,
Protein Engineering (1988), 2(3)209-218.

GRAY83:
Gray, WR, JE Rivier, R Galyean, LJ Cruz, and BM Olivera,
J Biol Chem, (1983), 258(20)12247-51.

GRAY84:
Gray, WR, FA Luque, R Galyean, E Atherton, and RC Sheppard, BL Stone, A Reyes, J Alford, M McIntosh, BM Olivera et al.
Biochemistry, (1984), 23(12)2796-802.

GUAR89:
Cuarino, A, R Giannella, and MR Thompson,
Infection and Immunity (1989), 57(2)649-52.

GUDM89:
Gudmundsdottir, A, PE Bell, MD Lundrigan, and C Bradbeer, and RJ Kadner,
J Bacteriol, (Dec 1989), 171(12)6526-33.

GUSS88:
Guss, JM, EA Merritt, RP Phizackerley, R Hedman, M Murata, KO Hodgson, HC Freeman,
Science (1988), 241:806-11.

GUZM89:
Guzman-Verduzco, L-M, and YM Kupersztoch,
Infection and Immunity (1989), 57(2)645-48.

GUZM90:
Guzman-Verduzco, L-M, and YM Kupersztoch,
Molec Microbiol (1990), 4:253-64.

HARD90:
Hard, T, E Kellenbach, R Boelens, BA Maler, K Dahlman, LP Freedman, J Carlstedt-Duke, KR Yamamoto, J-A Gustafsson, and R Kaptein,
Science (13 July 1990), 249:157-60.

HASH85:
Hashimoto, K, S Uchida, H Yoshida, Y Nishiuchi, S Sakakibara, and K Yukari,
Eur J Pharmacol (1985), 118(3)351-4.

HATA90:
Hatanaka, Y, E Yoshida, H Nakayama, and Y Kanaoka,
Chem Pharm Bull (Tokyo), (Jan 1990), 38:236-8.

HEDE89:
Hedegaard, L, and P Klemm,
Gene, (Dec 21 1989), 85(1)115-24.

HEIN87:
Heine, HG, J Kyngdon, and T Ferenci,
Gene (1987), 53:287-92.

HEIN88:
Heine, HG, G Francis, KS Lee, and T Ferenci,

J Bacteriol (April 1988), 170:1730-8.

HIDA90:
Hidaka, Y, K Sato, H Nakamura, J Kobayashi, Y Ohizumi, and Y SHimonishi,
FEBS Lett (1990), 264(1)29-32.

HOCJ85:
Ho, C, M Jasin, and P Schimmel,
Science (1985), 229:389-93.

HOLA89a:
Holak, TA, D Gondol, J Otlewski, and T Wilusz,
J Mol Biol (1989), 210:635-648.

HOLA89b:
Holak, TA, W Bode, R Huber, J Otlewski, and T Wilusz,
J Mol Biol (Dec 5 1989), 210(3)649-54.

HORV89:
Horvat, S, B Grgas, N Raos, and VI Simeon,
Int J Peptide Protein Res (1989), 34:346-51.

HOUG84:
Houghten, RA, JM Ostresh, and FA Klipstein,
Eur J Biochem (1984), 145:157-162.

ITOK79:
Ito, K, G Mandel, and W Wickner,
Proc Natl Acad Sci USA (1979), 76:1199-1203.

JANA89:
Janatova, J, KBM Reid, and AC Willis,
Biochem (1989), 28:4754-61.

JANI85:
Janin, J, and C Chothia,
Methods in Enzymology (1985), 115(28)420-430.

JENN89:
Jennings, PA, MM Bills, DO Irving, and JS Mattick,
Protein Eng, (Jan 1989), 2(5)365-9.

JUDD85:
Judd, RC,

Infect Immun (1985), 48(2)452-7.

JUDD86:
Judd, RC,
Infect Immun (1986), 54(2)408-14.

KATZ86:
Katz, BA, and A Kossiakoff,
J Biol Chem (1986), 261(33)15480-85.

KATZ90:
Katz, B, and AA Kossiakoff,
Proteins, Struct, Funct, and Genet (1990), 7:343-57.

KISH85:
Kishore, R, and P Balaram,
Biopolymers (1985), 24:2041-43.

KOBA89:
Kobayashi, Y, T Ohkubo, Y Kyogoku, Y Nishiuchi, S Sakakibara, W Braun, nad N Go,
Biochemistry (1989), 28:4853-60.

KUBO89:
Kubota, H, Y Hidaka, H Ozaki, H Ito, T Hirayama, Y Takeda, and Y Shimonishi,
Biochem Biophys Res Comm (1989), 161:229-235.

KUHN85a:
Kuhn, A, and W Wickner,
J Biol Chem (1985), 260:15914-15918.

KUHN85b:
Kuhn, A, and W Wickner,
J Biol Chem (1985), 260:15907-15913.

KUPE90:
Kupersztoch, YM, K Tachias, CR Moomaw, LA Dreyfus, R Urban, C Slaughter, and S Whipp,
J Bacteriol (1990), 172(5)2427-32.

LAM91:
Lam, KS, et al.
Nature (1991), 354:82-84.

LISS85:
Liss, LR, BL Johnson, and DB Oliver,
J Bacteriol (1985), 164(2)925-8.

LUND86:
Lundeen, M,
J Inorgan Biochem (1986), 27:151-62.

MACI88:
MacIntyre, S, R Freudl, ML Eschbach, and U Henning,
J Biol Chem (1988), 263(35)19053-9.

MANI82:
Maniatis, T, EF Fritsch, and J Sambrook,
Molecular Cloning,
Cold Spring Harbor Laboratory, 1982.

MANO86:
Manoil, C, and J Beckwith,
Science (1986), 233:1403-1408.

MANO88:
Manoil, C, D Boyd, and J Beckwith,
Topics in Genetics (1988), 4(8)223-6.

MARK86:
Marks, CB, M Vasser, P Ng, W Henzel, and S Anderson,
J Biol Chem (1986), 261:7115-7118.

MARK91:
Markland et al.
Gene (1991) 109:13-19.

MATS89:
Matsumura, M, WJ Becktel, M Levitt, and BW Matthews,
Proc Natl Acad Sci USA (1989), 86:6562-6.

MCCA90:
McCafferty, J, AD Griffiths, G Winter, and DJ Chiswell,
Nature, (6 Dec 1990), 348:552-4.

MCKE85:
McKern, NM, IJ O'Donnell, DJ Stewart, and BL Clark,
J Gen Microbiol (1985), 131(Pt 1)1-6.

MCWH89:
McWherter, CA, WF Walkenhorst, EJ Campbell, and GI Glover,
Biochemistry (1989), 28:5708-14.

MISR88a:
Misra, R, and SA Benson,
J Bacteriol (1988), 170(8)3611-7.

MISR88b:
Misra, R, and SA Benson,
J Bacteriol (1988), 170:528-33.

MORS87:
Morse, SA, TA Mietzner, G Bolen, A Le Faou, and G Schoolnik,
Antonie Van Leeuwenhoek (1987), 53(6)465-9.

MORS88:
Morse, SA, C-Y Chen, A LeFaou, and TA Meitzner,
Rev Infect Dis (1988), 10(Suppl 2)S306-10.

NICH88:
Nicholson, H, WJ Becktel, and BW MAtthews,
Nature (1988), 336:651-56.

NIKA84:
Nikaido, H, and HCP Wu,
Proc Natl Acad Sci USA (1984), 81:1048-52.

NISH82:
Nishiuchi, Y, and S Sakakibara,
FEBS Lett (1982), 148:260-2.

NISH86:
Nishiuchi, Y, K Kumagaye, Y Noda, TX Watanabe, and S Sakakibara,
Biopolymers, (1986), 25:S61-8.

OKAM87:
Okamoto, K, K Okamoto, J Yukitake, Y Kawamoto, and A Miyama,
Infection and Immunity (1987), 55:2121-2125.

OKAM88:
Okamoto, K, K Okamoto, J Yukitake, and A Miyama,
Infection and Immunity (1988), 56:2144-8.

OKAM90:
Okamoto, K, and M Takahara,
J Bacteriol (1990), 172(9)5260-65.

OLIP86:
Oliphant, AR, AL Nussbaum, and K Struhl,
Gene (1986), 44:177-183.

OLIP87:
Oliphant, AR, and K Struhl
Methods in Enzymology 155 (1987)568-582.
Editor Wu, R; Academic Press, New York.

OLIV87:
Olivera, BM, LJ Cruz, V de Santos, GW LeCheminant, D Griffin, R Zeikus, JM McIntosh, R Galyean, J Varga, WR Gray, et al.
Biochemistry, (1987), 26(8)2086-90.

OLIV90a:
Olivera, BM, J Rivier, C Clark, CA Ramilo, GP Corpuz, FC Abogadie, EE Mena, SR Woodward, DR Hillyard, LJ Cruz,
Science, (20 July 1990), 249:257-263.

ORND85:
Orndorff, PE, and S Falkow,
J Bacteriol (1985), 162:454-7.

OTLE87:
Otlewski, J, H Whatley, A Polanowski, and T Wilusz,
Biol Chem Hoppe-Seyler (1987), 368:1505-7.

PABO79:
Pabo, CO, RT Sauer, JM Sturtevant, and M Ptashne,
Proc Natl Acad Sci USA (1979), 76:1608-1612.

PABO86:

Pabo, CO, and EG Suchanek,
Biochem (1986), 25:5987-91.

PAGE88:
Pages, JM, and JM Bolla,
Eur J Biochem (1988), 176(3)655-60.

PAGE90:
Pages, JM, JM Bolla, A Bernadac, and D Fourel,
Biochimie (1990), 72:169-76.

PAKU86:
Pakula, AA, VB Young, and RT Sauer,
Proc Natl Acad Sci USA (1986), 83:8829-8833.

PANT87:
Pantoliano, MW, RC Ladner, PN Bryan, ML Rollence, JF Wood, and TL Poulos,
Biochem (1987), 26:2077-82.

PANT90:
Pantoliano, MW, and RC Ladner,
US Patent 4,908,773, 13 March 1990.

PARD89:
Pardi, A, A Galdes, J Florance, and D Maniconte,
Biochemistry (1989), 28:5494-5501.

PARG87:
Parge, HE, DE McRee, MA Capozza, SL Bernstein, ED Getzoff, and JA Tainer,
Antonie Van Leeuwenhoek (1987), 53(6)447-53.

PARM88:
Parmley, SF, and GP Smith,
Gene (1988), 73:305-318.

PARR88:
Parraga, G, SJ Horvath, A Eisen, WE Taylor, L Hood, ET Young, RE Klevit,
Science (1988), 241:1489-92.

PEAS88:
Pease, JHB, and DE Wemmer,
Biochem (1988), 27:8491-99.

PEAS90:
Pease, JHB, RW Storrs, and DE Wemmer,
Proc Natl Acad Sci USA (1990), 87:5643-47.

PERR84:
Perry, LJ, and R Wetzel,
Science (1984), 226:555-7.

PERR86:
Perry, LJ, and R Wetzel,
Biochem (1986), 25:733-39.

POTE83:
Poteete, AR,
J Mol Biol (1983), 171:401-418.

QUIO87:
Quiocho, FA, NK Vyas, JS Sack and MA Storey,
in Crystallography in Molecular Biology, Moras, D. et al., editors, Plenum Press, 1987.

RASC86:
Rasched, I, and E Oberer,
Microbiol Rev (1986) 50:401-427.

RASH84:
Rashin, A,
Biochemistry (1984), 23:5518.

REID88a:
Reidhaar-Olson, JF, and RT Sauer,
Science (1988), 241:53-57.

REID88b:
Reid, J, H Fung, K Gehring, PE Klebba, and H Nikaido,
J Biol Chem (1988), 263(16)7753-9.

RICH86:
Richards, JH,
Nature (1986), 323:187.

RIVI87b:
Rivier, J, R Galyean, WR Gray, A Azimi-Zonooz, JM McIntosh, LJ Cruz, and BM Olivera,
J Biol Chem, (1987), 262(3)1194-8.

ROBE86:
Roberts, S, and AR Rees
Protein Engineering (1986), 1:59-65.

ROSS81:
Rossman, M, and P Argos,
Ann Rev Biochem (1981), 50:497-532.

SALI88:
Sali, D, M Bycroft, and AR Fersht,
Nature (1988), 335:740-3.

SAUE86:
Sauer, RT, K Hehir, RS Stearman, MA Weiss, A Jeitler-Nilsson, EG Suchanek, and CO Pabo,
Biochem (1986), 25:5992-98.

SCHA78:
Schaller, H, E Beck, and M Takanami,
The Single-Stranded DNA Phages, Denhardt, D.T., D. Dressler, and D.S. Ray editors,
Cold Spring Harbor Laboratory, 1978., p139-163.

SCHN86:
Schnabel, E, W Schroeder, and G Reinhardt,
Biol Chem Hoppe-Seyler (1986), 367:1167-76.

SCHU79:
Schulz, GE, and RH Schirmer,
Principles of Protein Structure,
Springer-Verlag, New York, 1979.

SCOT87a:
Scott, MJ, CS Huckaby, I Kato, WJ Kohr, M Laskowski Jr., M-J Tsai and BW O'Malley,
J Biol Chem (1987), 262(12)5899-5907.

SCOT90:
Scott, JK, and GP Smith,
Science, (27 July 1990), 249:386-390.

SEKI85:
Sekizaki, T, H Akaski, and N Terakado,
Am J Vet Res (1985), 46:909-12.

SHIM87:
Shimonishi, Y, Y Hidaka, M Koizumi, M Hane, S Aimoto, T Takeda, T Miwatani, and Y Takeda,
FEBS Lett (1987), 215:165-170.

SILH77:
Silhavy, TJ, HA Shuman, J Beckwith, and M Schwartz,
Proc Natl Acad Sci USA (1977), 74(12)5411-5415.

SMIT85:
Smith GP,
Science (1985), 228:1315-1317.

SODE85:
Sodergren, EJ, J Davidson, RK Taylor, and TJ Silhavy,
J Bacteriol (1985), 162(3)1047-1053.

SOME85:
So, M, E Billyard, C Deal, E Getzoff, P Hagblom, TF Meyer, E Segal, and J Tainer,
Curr Top in Microbiol & Immunol (1985), 118:13-28.

STAD89:
Stader, J, LJ Gansheroff, and TJ Silhavy,
Genes & Develop (1989), 3:1045-1052.

STEI85:
Steiner,
BioScience Repts. (1985), 5:973ff.

SUMM91:
Summers, MF,
J Cell Biochem (1991) 45:41-8.

SUNX87:
Sun, XP, H Takeuchi, Y Okano, and Y Nozawa,
Comp Biochem Physiol [C], (1987), 87(2)363-6.

TAKA85:
Takao, T, N Tominaga, S Yoshimura, Y Shimonishi, S Hara, T Inoue, and A Miyama,
Eur J Biochem (1985), 152:199-206.

TAKE90:
Takeda, T, GB Nair, K Suzuki, and Y Shimonishi,
Infection and Immunity (1990), 58(9)2755-9.

TANK77:
Tan, NH, and ET Kaiser,
Biochemistry, (1977), 16:1531-41.

THOM85a:
Thompson, MR, M Luttrell, G Overmann, RA Giannella,
Analytical Biochem (1985), 148:26-36.

THOM85b:
Thompson, MR, and RA Giannella,
Infection & Immunity (1985), 47:834-36.

THOR88:
Thornton, JM, BL Sibinda, MS Edwards, and DJ Barlow,
BioEssays (?) SKG 3039 ??????

TOMM82:
Tommassen, J, P van der Ley, A van der Ende, H Bergmans, and B Lugtenberg,
Mol gen Genet (1982), 185:105-110.

TRIA88:
Trias, J, EY Rosenberg, and H Nikaido,
Biochim Biophys Acta (1988), 938:493-496.

VAND86:
van der Ley, P, M Struyve, and J Tommassen,
J Biol Chem (1986), 261(26)12222-5.

VAND90:
van der Werf, S, A Charbit, C Leclerc, V Mimic, J Ronco, M Girard, and M Hofnung,
Vaccine (1990), 8(3)269-77.

VERS86a:
Vershon, AK, K Blacker, and RT Sauer,
pp243-256 in Protein Engineering. Applications in Science, Medicine, and Industry,
Academic Press, 1986.

VERS86b:
Vershon, AK, JU Bowie, TM Karplus, and RT Sauer,
pp302-311 in Proteins: Structure, Function, and Genetics, Alan R. Liss, Inc., 1986.

VITA84:
Vita, C, D Dalzoppo, and A Fontana,
Biochemistry (1984), 23:5512-5519.

VOGE86:
Vogel, H, and F Jahnig,
J Mol Biol (1986), 190:191-99.

WATS87:
Molecular Biology of the Gene, Fourth Edition,
Watson, JD, NH Hopkins, JW Roberts, JA Steitz, and AM Weiner,
Benjamin/Cummings Publishing Company, Inc., Menlo Park, CA., 1987.

WEBS78:
Webster, RE, and JS Cashman,
The Single-Stranded DNA Phages, Denhardt, DT, D Dressler, and DS Ray editors, Cold Spring Harbor Laboratory, 1978., p557-569.

WEHM89:
Wehmeier, U, GA Sprenger, and JW Lengeler,
Mol Gen Genet (1989), 215(3)529-36.

WEIN83:
Weinstock, GM, C ap Rhys, ML Berman, B Hampar, D Jackson, TJ Silhavy, J Weisemann, and M Zweig,
Proc Natl Acad Sci USA (1983), 80:4432-4436.

WELL86:
Wells, JA, and DB Powers,
J Biol Chem (1986), 261:6564-70.

WEMM83:
Wemmer, D, and NR Kallenbach,
Biochem (1983), 22:1901-6.

WHAR86:
Wharton, RP,
The Binding Specificity Determinants of 434 Repressor.,
Harvard U. PhD Thesis, 1986,
University Microfilms, Ann Arbor, Michigan.

WIEC85:
Wieczorek, M, J Otlewski, J Cook, K Parks, J Leluk, A Wilimowska-Pelc, A Polanowski, T Wilusz, and L Laskowski, Jr,
Biochem Biophys Res Comm (1985), 126(2)646-652.

WILK84:
Wilkinson, AJ, AR Fersht, DM Blow, P Carter, and G Winter,
Nature (1984), 307:187-188.

WOOD90:
Woodward, SR, LJ Cruz, BM Olivera, and DR Hillyard,
EMBO J (1990), 9:1015-1020.

YOSH85:
Yoshimura, S, H Ikemura, H Watanabe, S Aimoto, Y Shimonishi, S Hara, T Takeda, T Miwatani, and Y Takeda,
FEBS Lett (1985), 181:138-42.

ZAFA88:
Zafaralla, GC, C Ramilo, WR Gray, R Karlstrom, BM Olivera, and LJ Cruz,
Biochemistry, (1988), 27(18)7102-5.

CLAIMS

1. In a process for developing novel binding proteins with a desired binding activity against a particular target material comprising providing a population of genetic packages, each displaying one or more copies of a particular potential binding domain as part of a chimeric outer surface protein thereof, said potential binding domain not being natively associated with the outer surface of said package, said population collectively displaying a plurality of different potential binding domains, the differentiation among said plurality of different potential binding domains occurring through the at least partially random variation of one or more predetermined amino acid positions, but not all amino acid positions, of said parental binding domain to randomly obtain at each said variable position an amino acid belonging to a predetermined set of two or more amino acids, the amino acids of said set occurring at said position in predetermined expected proportions; contacting the packages with the target material; and separating the packages according to their affinity for said target material;

the improvement comprising essentially each said potential binding domain being a mini-protein sequence of less than forty amino acids and having at least one intrachain covalent crosslink between at least a first amino acid position and a second amino acid position thereof, the amino acids at said first and second positions being invariant in all of the chimeric proteins displayed by said population, with those residues which participate in the formation of a covalent crosslink being invariant throughout said population, with the proviso that when the crosslink is in the form of a disulfide bond, the potential binding domain is a micro-protein sequence of less than forty amino acids.

2. The method of claim 1 wherein the crosslink is a disulfide bond and the the amino acids at the first and second amino acid positions are cysteines.

3. The method of claim 2 in which the micro-protein domain has a single disulfide bond and the span of the bond is not more than nine amino acid residues.

4. The method of claim 2 in which the micro-protein domain has a single disulfide bond, wherein the disulfide bond bridges a sequence of amino acids which under affinity separation conditions collectively assume a hairpin supersecondary structure.

5. The method of claim 4 wherein the hairpin secondary structure is selected from the group consisting of (a) an α helix, a turn, and a β strand; (b) an α helix, a turn, and an α helix; and (c) a β strand, a turn, and a β strand.

6. The method of claim 2 wherein the micro-protein domain comprises two intrachain disulfide bonds and preferably includes two clustered cysteines.

7. The method of claim 6 wherein the micro-protein domain has two disulfide bonds having a connectivity pattern of 1-3, 2-4.

8. The method of claim 2 wherein the micro-protein domain comprises three intrachain disulfide bonds and preferably includes two clustered cysteins.

9. The method of claim 8 wherein the micro-protein domain has three disulfide bonds having a connectivity pattern of 1-4, 2-5, 3-6.

10. The method of claim 7 wherein the micro-protein domain substantially corresponds in sequence to an α-conotoxin.

11. The method of claim 9 wherein the micro-protein domain substantially corresponds in sequence to a mu- or omega-conotoxin.

12. The method of claim 6 wherein the micro-protein domain substantially corresponds in sequence to a micro-protein selected from the group consisting of Escherichia coli heat

stable toxin I (ST_A), the bee venom apamin, or a squash-seed trypsin inhibitor, the scorpion toxin, charybdotoxin and secretory leukocyte protease inhibitor.

13. The method of claim 1 wherein the covalent crosslink includes a metal atom, such as zinc, iron, copper or cobalt.

14. The method of any of claims 1-13 wherein at least one variable amino acid position in said potential binding domains was encoded by a simply variegated codon selected from the group consisting of NNT, NNG, RNG, RMG, VNT, RRS, and SNT.

15. The method of any of claims 1-13 wherein none of the variable amino acid positions in said potential binding domain was encoded by a simply variegated codon selected from the group consisting of NNN, NNK and NNS.

16. The method of any of claims 1-13 wherein at least one variable amino acid position in said potential binding domains was encoded by a complexly variegated codon.

17. The method of any of claims 1-16 wherein the replicable genetic package is a phage, preferably a DNA phage other than phage lambda, more preferably a filamentous phage.

18. The method of claim 17 wherein the potential binding domain is fused with the major coat protein of a filamentous phage or a assemblable fragment thereof, or with the gene III protein of a filamentous phage or an assemblable fragment thereof.

19. The method of any of claims 1-16 wherein the replicable genetic package is a bacterial cell, such as strains of Escherichia coli, Salmonella typhimurium, Pseudomonas aeruginosa, Klebsiella pneumonia, Neisseria gonorrhoeae, or Bacillus subtilis, said DNA construct further comprises a periplasmic secretion signal sequence, and the potential binding domain is fused with a bacterial outer surface protein such as the lamB protein, OmpA, OmpC, OmpF, Phospholipase A, or pilin, or an assemblable segment thereof.

20. The method of any of claims 1-19 wherein said population is characterized by the display of at least 10^5 different potential binding domains, and wherein, for any potentially encoded potential binding domain, the probability that it will be displayed by at least one package in said population is at least 50%, more preferably at least 90%.

21. A library of display phage or cells, each displaying one or more copies of a particular potential binding domain as part of a chimeric outer surface protein thereof, said potential binding domain not being natively associated with the outer surface of said phage or cells, said library collectively displaying a plurality of different potential binding domains, the differentiation among said plurality of different potential binding domains occurring through the at least partially random variation of one or more predetermined amino acid positions, but not all amino acid positions, of said parental binding domain to randomly obtain at each said variable position an amino acid belonging to a predetermined set of two or more amino acids, the amino acids of said set occurring at said position in predetermined expected proportions,

essentially each said potential binding domain being a mini-protein sequence of less than sixty amino acids and having at least one intrachain covalent crosslink between at least a first amino acid position and a second amino acid position thereof, the amino acids at said first and second positions being invariant in all of the chimeric proteins displayed by said population, with those residues which participate in the formation of a covalent crosslink being invariant throughout said population, with the proviso that when the crosslink is a disulfide bond, the potential binding domain is a micro-protein of less than 40 residues.

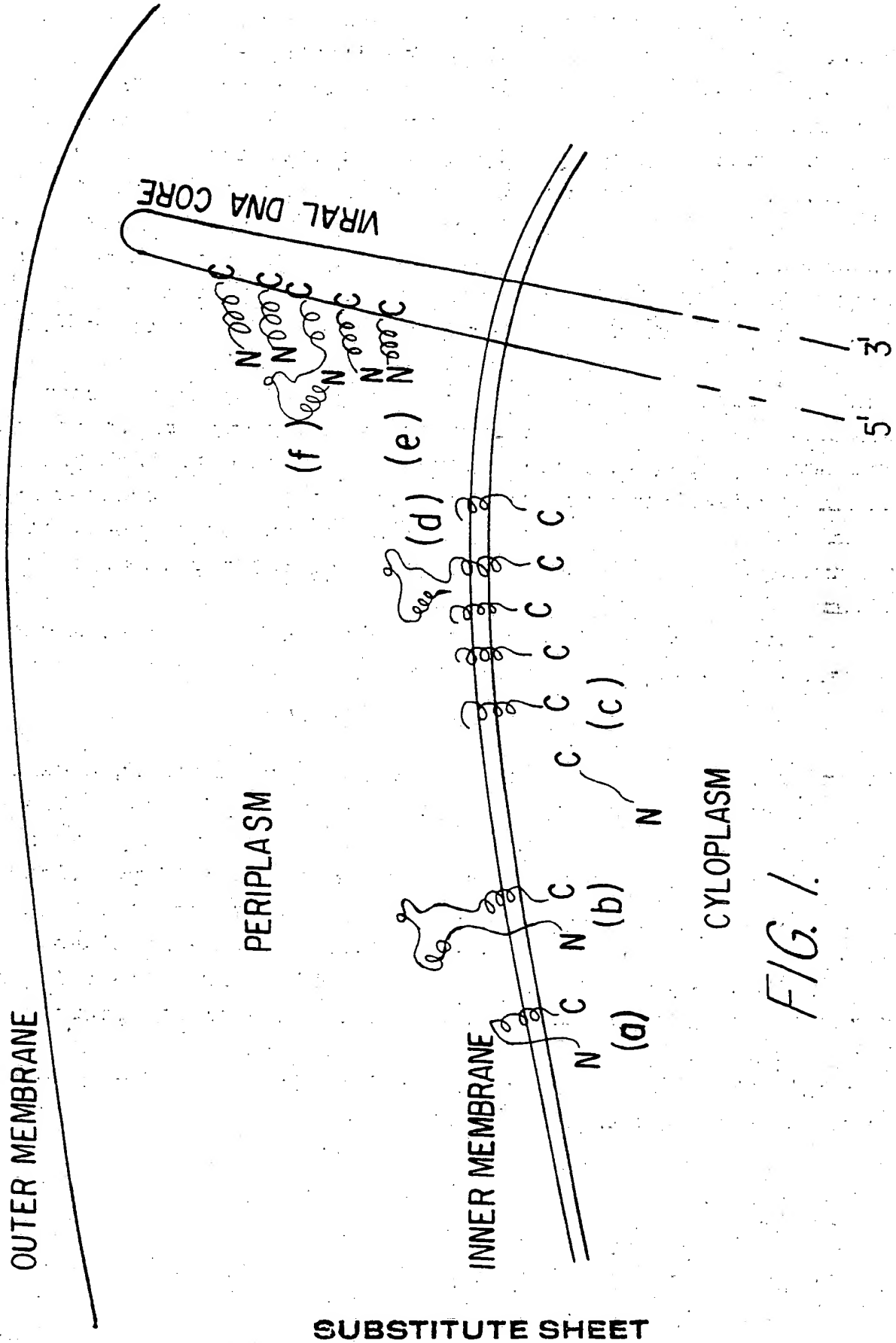

OUTER MEMBRANE
PERIPLASM
INNER MEMBRANE
CYLOPLASM
VIRAL DNA CORE
5'
3'
(a)
(b)
(c)
(d)
(e)
(f)
N
C

FIG. 1.

INTERNATIONAL SEARCH REPORT

PCT/US 92/01456

International Application No

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate)

According to International Patent Classification (IPC) or to both National Classification and IPC

Int.Cl. 5 C12N15/10; C12N15/62; C12N15/12; C12N15/63
C12P21/02; C07K13/00

II. FIELDS SEARCHED

Minimum Documentation Searched[7]

Classification System	Classification Symbols
Int.Cl. 5	C12N ; C12P ; C07K

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched[8]

III. DOCUMENTS CONSIDERED TO BE RELEVANT[9]

Category °	Citation of Document,[11] with indication, where appropriate, of the relevant passages[12]	Relevant to Claim No.[13]
Y	WO,A,9 002 809 (PROTEIN ENGINEERING CORPORATION) 22 March 1990 cited in the application see page 11, line 10 - page 14 see page 29, line 30 - page 48 see page 81, line 15 - page 102 see page 123, line 10 - page 137; claims ---	1-13, 17-21
Y	SCIENCE. vol. 249, 20 July 1990, LANCASTER, PA US pages 257 - 263; OLIVERA, B.M. ET AL.: 'Diversity of Conus neuropeptides' cited in the application see the whole document ---	1-11, 17-21

-/--

° Special categories of cited documents :[10]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report
29 JUNE 1992	15. 07. 92
International Searching Authority	**Signature of Authorized Officer**
EUROPEAN PATENT OFFICE	ANDRES S.M.

Form PCT/ISA/210 (second sheet) (January 1985)

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)

Category °	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claim No.
Y	SCIENCE. vol. 249, 13 July 1990, LANCASTER, PA US pages 157 - 160; HÄRD, T. ET AL.: 'Solution structure of the glucocorticoid receptor DNA-binding domain' cited in the application see the whole document	13
Y	BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS. vol. 161, no. 1, 30 May 1989, DULUTH, MINNESOTA US pages 229 - 235; KUBOTA, H. ET AL.: 'A long-acting heat-stable enterotoxin analog of enterotoxigenic Escherichia coli with a single D-amino acid' cited in the application see the whole document	12
A	SCIENCE. vol. 249, 27 July 1990, LANCASTER, PA US pages 404 - 406; DEVLIN, J.J. ET AL.: 'Random peptide libraries: a source of specific protein binding molecules' cited in the application see the whole document	1-21
A	SCIENCE. vol. 249, 27 July 1990, LANCASTER, PA US pages 386 - 390; SCOTT, J.K. & SMITH, G.P.: 'Searching for peptide ligands with an epitope library' cited in the application see the whole document	1-21
A	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA. vol. 87, 1990, WASHINGTON US pages 6378 - 6382; CWIRLA, S.E. ET AL.: 'Peptides on a phage: a vast library of peptides for identifying ligands' cited in the application see the whole document	1-21

ANNEX TO THE INTERNATIONAL SEARCH REPORT ON INTERNATIONAL PATENT APPLICATION NO. US 9201456 SA 58102

This annex lists the patent family members relating to the patent documents cited in the above-mentioned international search report. The members are as contained in the European Patent Office EDP file on 29/06/92
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO-A-9002809	22-03-90	AU-A- 4308689	02-04-90
		EP-A- 0436597	17-07-91